A Novel Method for Modeling Hierarchical Developmental Toxicity Data and Calculating Joint Risk BMDs Based on the Plackett-Dale Distribution

Dissertation Advisor: Paul Catalano Frederick Prichard Cudhea

A Novel Method for Modeling Hierarchical Developmental Toxicity Data and Calculating Joint Risk BMDs Based on the Plackett-Dale Distribution

Abstract

In developmental toxicity studies, multiple levels of correlation exist between multiple outcomes of interest, complicating the estimation of models and risk assessment for data collected from these studies. The first chapter describes these multiple layers of correlation, the problems that arise from them, and provides a detailed literature review of the statistical methodology developed in order to address these problems.

The second chapter presents a method for modeling death and malformation outcomes based on the bivariate Plackett-Dale distribution. The method defines three association parameters to describe all litter-level correlations, and then derives bivariate Plackett-Dale distributions based on these three associations. A pseudolikelihood based on the probability mass functions of these distributions is used as a basis for estimating the model parameters for death and malformation as well as the three association parameters. The model relaxes the conditional independence assumption and, unlike methods assuming an underlying latent normal distribution, allows for assuming death and malformation following a Bernoulli distribution. The method is applied to two different datasets and compared to other methods. The third chapter examines the small sample behavior of the proposed model in chapter two with simulations. A comparison to other methods, as well as an examination of the robustness of the model to misspecification of second order parameters is also presented.

The fourth chapter proposes a method for joint risk estimation for the proposed

model as well as for Carey's method, an approach to modeling the data without assuming conditional independence but with no development for joint risk estimation. These methods were compared to methods that do assume conditional independence with the two data sets used in chapter two, as well as with simulations.

Finally, the fifth chapter summarizes the work presented and its specific contribution to the field of analysis of developmental toxicity data. The advantages and limitations of the proposed model are discussed, as well as and possible avenues for future research.

Contents

Acknowledgments

I would like to thank, first and foremost, my advisor, Paul Catalano, for his pivotal guidance. His mentorship, patience, and insights in all aspects of this project have been essential. From how to think about the problem conceptually, to resolving computational difficulties, to how to best present these ideas with clarity, he has helped overcome all road blocks I faced as I worked on this project. His generous spirit made him a pleasure to work with. I would also like to thank my committee members, Brent Coull and Francesca Dominici, for their valuable insights. The fresh pairs of eyes they provided throughout this process made the thesis a much stronger work.

I would also like to thank my fellow students, as well as staff, current and former, for letting me pick their brains, not just for statistical knowledge, but for life knowledge as well. And many thanks to my officemates for providing, at times annoying, but ultimately necessary distractions throughout the day. I am especially grateful to those who gave me encouraging words towards the end when I needed to hear them the most.

Finally, I would like to thank my family for their unceasing support during my time here. I'd like to especially thank my mother, Fukiko Cudhea, for her nurturing care in times of stress. Without her support, before and during my stint as a Ph.D student, I never would have finished.

A Review of Methods for Modeling Developmental Toxicology Outcomes

Frederick Prichard Cudhea

Department of Biostatistics
Harvard School of Public Health

1.1 Introduction

Controlled animal studies play an important role in determining safe doses for environmental toxic substances, drugs, and other chemical agents. Unlike studies that determine efficacy of treatments, studies that determine the toxicity of a substance cannot use human subjects. Thus, animal studies are often the only method where experimental data can be obtained to assess toxicity. Along with other information, a proper analysis of experimental data can be helpful to regulators who need to decide on an acceptable dose of the substance in question. These assays can range from studies for determining risk of various types of cancer, an area where much of the early work in dose-response modeling and risk assessment was developed, to studies for various non-cancer related toxicities such as developmental and neurotoxicities.

In developmental toxicology studies, female animals (rodents or rabbits) are mated, and then exposed to specific doses of the toxin under study. Although sample sizes may vary by study, it is recommended at least 20 pregnant animals (dams) are assigned to each dose group, and the study have at least three dose-groups in addition to a control group (Kimmel and Price, 1990) (United States Environmental Protection Agency, 1991). During gestation some embryos are resorbed back into the uterus while others mature but do not survive gestation. These two embryolethality outcomes are known as resorptions and fetal deaths, respectively. Fetuses that survive gestation are at risk for adverse events including low birth weight and skeletal, visceral, and external malformations. Before natural birth, the dams are sacrificed and uterine contents are examined. The main endpoints of interest are typically the number of embryolethalities, and for fetuses that would be born, number of malformed fetuses (along with what kind of malformation) and fetal weights and lengths. Table 1.1 exhibits summary data by dose from a developmental toxicity data of Ethylene Glycol (EG) in rats reported by Price, Kimmel, Tyl, and Marr (Price et al., 1985).

The trends observed in this study are consistent with what is seen in many develop-

Table 1.1: Summary Results from a Developmental Toxicity Study of EG in Rats

Dose (g/kg)	Dams	Number of Implants Mean (SD)	Non-Live* (%)	Litter Size Mean (SD)	Malformation† No. (%)	Weight (g) Mean (SD)‡
0	28	14.21 (.26)	4.70	13.54 (.28)	5 (1.37)	3.404 (.052)
1.25	28	13.64 (.33)	6.35	12.75 (.38)	21 (6.65)	3.312 (.058)
2.50	29	12.72 (.62)	6.27	11.90 (.60)	86 (25.11)	2.916 (.056)
5.00	27	13.44 (.56)	21.34	11.04 (.79)	197 (75.43)	2.388 (.089)

∗ Number of dead or resorbed fetuses
† Number of fetuses with at least one malformation
‡ Ignoring clustering

mental toxicology assays. As is shown in Table 1.1, as dosage increases, the number of implants tends to remain the same, due to dose being assigned after mating, while the embryolethality rate tends to increase. Accordingly, litter size tends to decrease as dose increases. Note that the standard deviation of litter size also tends to increase with dose, illustrating litter response heterogeneity. For the live outcomes, malformation rate increases while fetal weight decreases as dose increases. Note, also, that at the highest dose of the study, standard deviation is significantly larger for fetal weights, again indicating more heterogeneity at the higheset dose level.

Developmental toxicity data can have complexities that make proper analysis challenging. First, the observed endpoints are clustered into litters, and animals from the same litters tend to be correlated. Second, among living fetuses, different outcomes from the same fetus may also be correlated. Third, live outcomes, such as fetal weight and malformations, have a hierarchical relationship with death, which further complicates the interpretation of the data. Calculating a proper safe dose needs to involve taking into account both intra-litter and inter-outcome correlations, as well as the hierarchical relationship between live fetal outcomes and number of dead fetuses in a given litter. Figure 4.1 shows the relationships between the various commonly measured outcomes in developmental toxicity.

1.2 Risk Assessment

Historically, the NOAEL (No Observed Adverse Effect Level) played an important role in determining safes doses of toxins (Catalano and Ryan, 1994). The NOAEL is defined as the largest dose in a toxicology experiment in which no statistically significant adverse effect is observed. The NOAEL has several weaknesses that make it unattractive for developmental toxicology. First, NOAEL studies are restricted to actual doses from the experiment, so a NOAEL may not even exist for a particular study. Second, the NOAEL approach to finding safe doses encourages small sample size studies. Since the NOAEL is determined by testing for a difference between adverse effects in the control group and

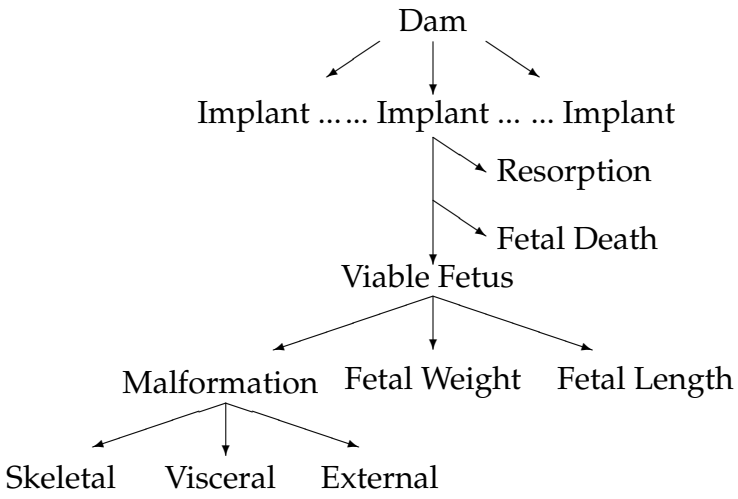

Figure 1.1: Outcomes in Developmental Toxicity

the dose groups, small sample studies with low power are less likely to detect a statistically significant difference at lower dose groups, leading, paradoxically, to higher acceptable doses. Perhaps the most damning weakness for the NOAEL approach is it does not provide any corresponding estimate of associated risk (Crump, 1984).

For these reasons, the NOAEL has been mostly abandoned in favor of the Benchmark Dose (BMD) method, first proposed by Crump (Crump, 1984). Instead of using an ANOVA-like approach where hypothesis tests are the major tool for determining safe doses, Crump proposes fitting a quantitative, continuous dose-response model to the experimental data, for example, in the form of $p(dose) = f(\beta_0 + \beta_1 dose)$ where f is a link function that ensures $p(dose)$ is bounded between 0 and 1. Popular dose-response models that have been used in developmental toxicology studies include the probit model, the logit model, extreme-value model, and Weibull model:

Probit:	$p(dose) = \Phi(\beta_0 + \beta_1 dose)$
Logit :	$p(dose) = \dfrac{1}{1 + exp[-(\beta_0 + \beta_1 dose)]}$
Extreme-Value :	$p(dose) = 1 - exp[-exp(\beta_0 + \beta_1 dose)]$
Weibull :	$p(dose) = 1 - exp[-(\beta_0 + \beta_1 dose^{\gamma})]$

When working with binary outcomes such as malformations or fetal deaths, the BMD is defined as the dose that corresponds to a pre-specified risk above background, known as the Benchmark Dose (BMR). For example, if we were interested in an extra risk (BMR) of 0.1, the BMD is the dose that solves $r(dose) = \frac{p(dose)-p(0)}{1-p(0)} = 0.1$, where $p(dose)$ is the probability of an adverse event at the specified dose, defined by the dose-response model. Alternatively, one can define the BMD using an additional risk function, in which case the BMD solves $p(dose) - p(0) = 0.1$. Extra risk is sometimes preferred over additional risk because, extra risk can be interpreted as the added risk of an adverse event out of those not affected had they not received a dose.

Toxicologists are often also interested in continuous outcomes such as fetal weight. For these endpoints, determining what outcomes constitute an adverse event is not as clear as for binary endpoints. In this situation, a cut-off value is usually determined based on the estimated mean and standard deviation of the control group. Given that, in the case of fetal weight, an abnormally low outcome is cause for concern, the cut-off of point, w_c can follow the formula: $w_c = w(0) - k \times sd(0)$ where $sd(0)$ is the estimated standard deviation for the control group and k is an arbitrary chosen threshold, generally ranging from 1.5 to 3. Alternatively, a cutoff can be determined by a low percentile (for example, 1%) of the control group distribution. Once the cut-off is determined, probability of an adverse event is defined as $p(w < w_c)$ (Gaylor et al., 1998).

Because a relationship betwen the dose and outcome is specified, a BMD should exist in every experiment and is not limited to the experimental doses. More importantly, a confidence interval can be placed on the BMD so that a 95% lower bound, called the $\text{BMDL}_{.95}$, can be calculated. Because a safe dose will be determined by the BMD's lower bound, this approach encourages larger sample size studies.

There are multiple methods for calculating the $\text{BMDL}_{.95}$. One intuitive approach is to let $BMDL_{.95} = BMD - 1.645\sqrt{var(BMD)}$. However, this approach allows for nonsensical negative BMDLs. Kimmel and Gaylor propose calculating the dose that corresponds an excess risk of 0.1 for the 95% upper confidence bound of the dose-

response curve (Kimmel and Gaylor, 1988). This translates to finding the dose that solves $\hat{r}(dose) + 1.645se(\hat{r}(dose)) = 0.1$, where $se(\hat{r}(dose))$ is obtained using the delta method. However this method is not invariant under parameter transformations. Another method, preferred by several authors, is to calculate confidence bounds based on likelihood ratio statistics. Let l_{max} be the unrestricted maximized log-likelihood and l_1 the log-likelihood under some constraint. The dose that satisfies $2(l_{max} - l_1) = 1.645^2$ and minimizes the BMD is a $BMDL_{.95}$. This method is preferred since it is invariant under transformations. However, it can only be used when a full likelihood distribution is assumed.

The methods described above only use information from a single adverse outcome to determine the safe dose. However, it is also important to consider the joint risk from multiple outcomes. Traditionally, when multiple outcomes are of interest, BMDs or NOAELs are calculated for each outcome and the smallest of these doses is chosen as the safe dose. In the case where adverse outcome are highly correlated, this approach is reasonable. However, in general, this approach is not satisfactory. For example, consider a case where there are two outcomes of interest, malformation and fetal weight. If two separate models are fit and two BMDs are calculated, one for each outcome, and the malformation BMD is chosen as the "more conservative" BMD, then the BMD may not take into account the additional toxic effects the substance poses to fetal weight. If the two outcomes are highly correlated, this additional toxic effect may be small. However, if the two outcomes are nearly independent, this additional toxic effect may be quite large, and ignoring this effect will lead to underestimating the safe dose (Ryan, 1992).

A more accurate approach would be to calculate a BMD based on the combined risk of all outcomes, while still allowing separate descriptions of dose-response relationships for each outcome. That is, instead of calculating a BMD for each outcome and choosing the most conservative, it is more advantageous to be able calculate one BMD where $p(dose)$ is the probability of any adverse outcome at the specified dose (for example, the probability of malformation or low fetal weight). Note that, in order to have a formula

for the joint probability, $p(dose)$, it is necessary to clearly define the correlations of the multiples responses, in some cases within and between animals.

1.3 Methods for Accommodating Litter Effects

The litter effect is an ever present issue in the analysis of developmental toxicology data and much of the early statistical work in this area focused on this problem. Outcomes from fetuses of the same dam tend to be correlated, and this correlation must be taken into account for a valid analysis. An analysis ignoring the litter effect will tend to underestimate variances, and thus, lead to misleadingly low p-values. Willliams proposed a famous model in which fetuses from the same dam share the same probability of malformation, but malformation probability would differ by dam (Williams, 1975). For dam k, let n_k be the number of live fetuses and M_k be the number of malformations. The so called beta-binomial model is a hierarchical model where $M_k|n_k, p_k \sim Binomial(n_k, p_k)$ and $p_k \sim Beta(\alpha(dose), \gamma(dose))$, so

$$P(M_k = y|n_k) = \binom{M_k}{y} \frac{B(\alpha(dose) + y, n_k + \gamma(dose) - y)}{B(\alpha(dose), \gamma(dose))}$$

The model assumes that intra-litter correlation is always positive. In the setting of developmental toxicity, it is expected that fetuses from the same litter will have similar outcomes, so this is considered a reasonable assumption.

For the purposes of parameter estimation, it is advantageous to reparameterize the model parameters to $\mu(dose) = \frac{\alpha(dose)}{\alpha(dose)+\gamma(dose)}$ and $\theta(dose) = \frac{1}{\alpha(dose)+\gamma(dose)}$. Here, $E[M_k] = n_k\mu(dose)$ and $Var[M_k] = n_k\mu(dose)(1 - \mu(dose))(1 + (n_k - 1)\frac{\theta(dose)}{1+\theta(dose)})$. Under this reparameterization, the intra-litter correlation for litter k is $\rho_m(dose) = \frac{\theta(dose)}{1+\theta(dose)}$. Note that $(1 + (n_k - 1)\frac{\theta}{1+\theta})$ is an inflation factor to the binomial variance which takes into account the extra-variation induced by the intra-litter correlation.

Also, under this reparameterization, the probability of observing a malformation for a fetus in litter k is $\mu(dose)$. While it is possible to estimate μ separately for each dose group,

a more useful approach, in terms of risk assessment, is to fit a dose-response model by letting $\mu = f(\beta_0 + \beta_1 dose)$. Then, through maximum likelihood estimation, β_0 and β_1 (as well as $\theta(dose)$) can be estimated to calculate a BMD and its associated BMDL.

Other notable extensions to the binomial model include the correlated binomial model (Kupper and Haseman, 1978) which allowed for negative intra-litter correlation, and the multiplicative binomial model (Altham, 1978). The correlated binomial model assumes that the correlation parameter is bounded, and these bounds are a function of n_k and p_k. Indeed, in the case of binary data, it is impossible to have a correlation of -1, except for the special case when $n = 2$. A perfect negative correlation implies one observed malformation in a litter implies all other fetuses will not be malformed, while one observed non-malformation corresponds to all other observations being malformations. This logic is contradictory in the case where litter size is greater than 2.

Many other extensions to handle correlated binary data exist but only a handful of important models will be described in this section. Ochi and Prentice, instead of generalizing the binomial model, took the approach of using the multivariate normal distribution, exploiting its flexible correlation structure (Ochi and Prentice, 1984). They assume that malformations from litter k are determined by latent variables, $\tilde{\mathbf{m}}_k$, that follow a multivariate normal distribution with mean $\mu_k \mathbf{1}_n$ and variance-covariance matrix $\sigma_k^2((1-\rho_k)\mathbf{I}_{n_k} + \rho_k \mathbf{J}_{n_k})$ where σ_k^2 is the variance of the latent variable and ρ_k is the common intra-litter correlation. Without loss of generality, the threshold that $\tilde{m}_{jk}$ needs to surpass for a malformation to be observed is assigned to be 0.

These assumptions lead to the correlated-probit model among the observable outcomes:

$$P(M_k = y) = \binom{n_k}{y} \int_A \phi_n(\tilde{\mathbf{m}}_k | \mu_k, \sigma_k, \rho_k)\, d\tilde{\mathbf{m}}_k$$

where $A = (\mathbf{m}_k | (\tilde{m}_{jk} > 0, j \leq y) \cup (\tilde{m}_{jk} \leq 0, j > y))$.
The area of integration, A, reflects that in order to observe y malformations, y of n_k latent variables must exceed 0 and the remaining $n_k - y$ latent variables must be less than y. Note that the above likelihood formulation contains an n_k dimensional integral, making it dif-

ficult to compute the first and second derivatives of the log likelihood necessary for estimating the regression parameters. Ochi and Prentice re-express the log likelihood derivatives in a form where the approximation of Mendell and Elston can be used (Mendell and Elston, 1974). Also, note that the correlation parameter in this model does not have additional constraints like the correlated-binomial model, because the model is based on the continuous normal distribution.

The Ochi-Prentice model is an important contribution to developmental toxicology methodology in that it introduces the concept of using latent variables to model binary outcomes. For many toxicologists, the concept of a quantal outcome being defined by whether an occult latent variable crosses a threshold is attractive from a biological theory perspective. However, given the computation complexity of the model and the the existence of simpler, more intuitive models based on the binomial distribution, the latent variable approach may seem unnecessary. Yet, the use of latent variables becomes a common feature in later models that incorporate both binary and continuous outcomes, since the latent formulation of a binary outcome can serve as a link between binary and continuous outcomes. These models are discussed in more detail in the mixed outcome section of the paper.

Rai and Van Ryzin (Rai and Van Ryzin, 1985), instead of developing a model based on classical statistical distributions, attempted to develop a more biologically motivated model based on the one-hit dose-response model, a concept borrowed from early carcinogenicity studies where the one-hit model was an established and popular method for cancer modeling and low-dose extrapolation. Biologically, the model assumes that only one "hit" or one genetic mutation from a toxic insult is necessary to begin the cascade to change a normal cell to a cancer cell. The dose-response model Rai and Van Ryzin propose is as follows:

$$P(m_{jk} = 1|dose, n_k) = (1 - e^{-(\beta_0+\beta_1 dose)})e^{(-n_k(\theta_0+\theta_1 dose))}$$

where $dose \geq 0, n_k \geq 0, \beta_0 \geq 0, \beta_1 \geq 0$ and $\theta_0 + \theta_1 dose \geq 0$ for all d_k. The first factor, $(1 - e^{-(\beta_0+\beta_1 dose)})$, can be interpreted as the probability of a toxic event occurring in dam

k at the specified dose, such that its offspring may be affected as well, and the second factor, $e^{(-n_k(\theta_0+\theta_1 dose))}$, can be interpreted as the conditional probability of a fetus from dam k experiencing an adverse event given that dam k experienced a toxic event where the litter size for dam k is n_k. Instead of taking into account the litter effect through the likelihood, Rai and Van Ryzin include litter size as part of the dose-response model as an ad-hoc method for handling the litter effect. Unfortunately, others have shown that this method does not account for all of the extra-binomial variation inherent in the data and, in general, does not fit the observed data well (Carr and Portier, 1991). Because the approach of fitting biologically motivated models have not successfully produced good fitting models, further research in developmental toxicity models is motivated less from biological theory and more on statistical flexibility.

All of the methods mentioned above specify a likelihood model for the outcome. Thus if the assumed likelihood model is misspecified, resulting parameters can be biased. Liang and Zeger's generalized estimating equations (GEE) (Liang and Zeger, 1986), an extension of the quasi-likelihood method (Wedderburn, 1974), makes it possible to estimate regression model parameters without having to correctly specify the distributions and correlations of the various outcomes. For this reason, the GEE has been a popular alternative to likelihood methods in many areas, including non-cancer toxicology studies.

With GEEs, the following estimating equations:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{k=1}^{K} \mathbf{D}_k^T \mathbf{V}_k^{-1}(\mathbf{m}_k - \boldsymbol{\mu}_k) = 0$$

are solved for $\boldsymbol{\beta}$, where $\mathbf{m}_k$ is the outcome vector (in this case, say, malformation) for litter k, $\boldsymbol{\mu}_k$ is the mean vector for the outcomes (in this case, the vector of malformation probabilities $p(dose_k)\mathbf{1}_{n_k}$), $\mathbf{D}_k = \frac{\partial \boldsymbol{\mu}_{jk}}{\partial \boldsymbol{\beta}}$, and $\mathbf{V}_k = \mathbf{A}_k^{\frac{1}{2}}\mathbf{R}(\alpha)\mathbf{A}_k^{\frac{1}{2}}$ is the working covariance matrix for $\mathbf{m}_k$ where $\mathbf{A}_k = diag[var(m_{jk})]$, $\mathbf{R}(\alpha)$ is the assumed correlation matrix for $\mathbf{m}_{jk}$ and α is the parameter characterizing the correlation. When working with binary outcomes such as malformations, $var(m_{jk}) = p(dose_k)(1 - p(dose_k))]$. In developmental toxicity, where it is reasonable to assume fetuses within a litter are equally correlated, a compound symmetry structure is usually chosen for $\mathbf{R}(\alpha)$. That is, $\mathbf{R}(\alpha) = (1-\alpha)\mathbf{I} + \alpha\mathbf{J}$.

If this is the correct correlation structure, then the covariance estimator for $\hat{\beta}$ is $\Sigma_{\hat{\beta},N} = \left(\sum_{k=1}^{K} \mathbf{D}_k^T \mathbf{V}_k^{-1} \mathbf{D}_k\right)^{-1}$. However, one can also use the sandwich estimator:

$$\Sigma_{\hat{\beta},R} = \left(\sum_{k=1}^{K} \mathbf{D}_k^T \mathbf{V}_k^{-1} \mathbf{D}_k\right)^{-1} \times \left(\sum_{k=1}^{K} \mathbf{D}_k^T \mathbf{V}_k^{-1} (\mathbf{m}_k - \boldsymbol{\mu}_k)(\mathbf{m}_k - \boldsymbol{\mu}_k)^T \mathbf{V}_k^{-1} \mathbf{D}_k\right) \left(\sum_{k=1}^{K} \mathbf{D}_k^T \mathbf{V}_k^{-1} \mathbf{D}_k\right)^{-1} \tag{1.1}$$

which is a consistent estimator, regardless of whether $\mathbf{R}(\alpha)$ is correctly specified.

Like the likelihood methods discussed above (not including the Rai-Van Ryzin model), there is a great degree of flexibility in terms of choosing a dose-response model. In general, one can fit any model that follows the form: $\mu_k = p(\mathbf{m}_k) = f(\mathbf{X}\beta)$, where $\mathbf{X}$ is the $n_k \times p$ matrix of covariates for $\mathbf{m}_k$ (including, of course, dose) and $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression parameters. Note that the likelihood models described above are all litter-level models since they treat the number of malformations as the outcome of interest. Thus, if one were to add additional covariates to the dose-response model, the new covariates would also have to be at the litter-level. By fitting a mean model where the outcomes are the binary malformation status of each fetus instead of the malformation count, we are free to allow fetus-specific covariates, such as sex, to be included in the model. Furthermore, the GEE approach using the sandwich estimator is a popular method, since it is robust to variance misspecification and can include a wider range of covariates.

However, the fact that minimal assumptions need to be made to use GEEs also poses a slight disadvantage in terms of quantitative risk assessment. Specifically, BMD calculations that require the likelihood cannot be completed since no likelihood is specified in this approach.

1.4 Multiple Binary Outcomes

In non-cancer studies, toxicologists are almost always interested in more than one type of adverse outcome. For example, there may be interest in probability of death in addition to probability of malformation, or in different kinds of malformations (e.g. skeletal, visceral, and external). The possible inter-outcome correlations further complicates analyses in this scenario.

Ryan and Lefkopoulou proposed a robust method, using GEEs, to model multiple binary outcomes (Lefkpoulou et al., 1989). Letting $M_{i\cdot k}$ be the number of malformations of type i from dam k and p_{ik} be the probability of fetus from litter k has a malformation of type i, they specify the moments for $M_{i\cdot k}$ as $E[M_{i\cdot k}] = n_k p_{ik}$ and $Var[M_{i\cdot k}] = n_k p_{ik}(1-p_{ik})\phi$ where ϕ is a dispersion parameter which takes into account the extra-binomial variation due to the litter effect. Then the following GEEs are used to solve for the logistic regression model $logit(p_{ik}) = \delta_i + \mathbf{X}_k\boldsymbol{\beta}$:

$$\mathbf{U}(\delta_\mathbf{i}, \boldsymbol{\beta}) = \sum_{k=1}^{K} \mathbf{D_k^T V_k^{-1}(M_k - n_k p_k)} = 0$$

where $\mathbf{D}_k^T$ is the matrix of mean derivatives, $\frac{\partial \mathbf{p}_k}{\partial \boldsymbol{\beta}_k}$, $\mathbf{M}_k$ is the vector of counts for each type of outcome, $\mathbf{p}_k$ is the vector of probabilities for each type of outcome, n_k is the the number of fetuses in dam k and K is the number of dams in the study. $\mathbf{V_k} = N_k\phi\mathbf{A}_k^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{A}_k^{1/2}$ where $\mathbf{A}_k = diag[p_{ik}(1-p_{ik})]$ and $\mathbf{R}(\boldsymbol{\alpha})$ is the matrix characterizing the correlation between the multiple outcomes within one fetus. Note that the variance matrix contains two parameters to account for two different types of correlation: α, which describes the intra-fetus correlation between the various outcomes, and ϕ, the dispersion parameter for $Var[M_{i\cdot k}]$, which is related to the correlation of two fetuses in the same litter on the same outcome ($\rho_k = \frac{\phi-1}{n_k-1}$).

Note that in this particular mean model, there is a unique δ_i for each response, but only a single β. Thus, the model assumes that the dose-response curves for each response are parallel, but with different intercepts. In other words, this mean model allows for

the estimation of one parameter to describe the common dose effect of across all outcomes. This assumption, that dose affects all outcomes in the same way, makes testing for an overall dose effect relatively simple since there is only a singe degree of freedom used for estimating the dose effect. However, for some data sets, this assumption may be too restrictive and a more flexible model such as $logit(p_{ik}) = \alpha_i + \mathbf{X}_k\boldsymbol{\beta}_i$, where dose is assumed to have a separate effect on each outcome, may be more appropriate. Indeed, being able to separately quantify the dose-response for each outcome is of interest to many toxicologists, especially since it is possible that some outcomes may be significantly more sensitive to dose than others. In cases where only a small proportion of the outcomes are sensitive to the toxin, an analysis assuming a common dose effect may misleadingly conclude that the toxin has no overall effect when in fact some outcomes of interest are affected by the toxin while others are unaffected. Also note that, unlike the GEE approach described earlier, this method treats the outcome of interest as malformation counts. Thus, including fetus-level covariates is not possible.

Using GEEs to handle multiple outcomes and the litter effect is advantageous since the method gives robust parameter estimates that are not biased from mis-specifying the covariance matrix, and minimal assumptions are required. However, because this method does not specify how to characterize the joint probability of multiple outcomes, it cannot be used to calculate a joint BMD, even though the correlation parameters can be easily estimated through method of moments.

1.5 Mixed Outcomes

The way toxic substances negatively affect fetal development is not limited to causing deaths and malformations. A low fetal weight from a fetus without the presence of a malformation can still indicate that a substance has a harmful effect and may be a sensitive outcome of toxicity, especially at low doses. Thus, in the interest of utilizing all available information to assess risk, toxicologists are also interested in fetal weights. However, the fact that fetal weight is a continuous outcome, and known to be correlated with binary

outcomes like malformation status, presents an analytic complication (Ryan et al., 1991). The simplest approach to this problem is to dichotomize the continuous outcomes by choosing an arbitrary cutpoint and defining an outcome below that cutpoint as an adverse event. This allows the use of methods developed for multiple binary outcomes. However, this approach essentially ignores that fetal weight is measured on a continuous scale and thus leads to a loss of information, both in terms of statistical efficiency and in terms of losing the ability to quantify how dose directly affects fetal weight. Thus, the models presented in this section maintain fetal weight as a continuous outcome while still accounting for the litter effect and inter-outcome correlation.

A common feature of the models described below is that they all circumvent defining the joint probability of malformation and fetal weight. Instead, they rely on the fact that the joint likelihood can be expressed as the product of the marginal distribution of one variable and the conditional distribution of the other variable. For example, the joint density of fetal weight and malformation, $f_{m,w}(m, w)$, can be expressed as $f_{m|w}(m|w)f_w(w)$, the product of conditional probability of malformation given weight and the marginal density of weight. Using this factorization allows for separately modeling mean and malformation based on simpler likelihood models while also taking into account that the two outcome are correlated.

One such model, proposed by Catalano and Ryan (Catalano and Ryan, 1992) assumes the underlying distribution for the outcomes is a multivariate normal, much like the Ochi-Prentice model. Malformation, the observed binary outcome, is assumed to be determined by a latent variable $\tilde{m}$ which follows the normal distribution. By taking this latent variable approach, it is possible to characterize the intra-outcome and inter-outcome correlations between these types of outcomes. More specifically, the proposed model is:

$$w_{jk} = \alpha_0 + \alpha_1 d_k + \epsilon_{wjk}$$
$$\tilde{m}_{jk} = \beta_0 + \beta_1 d_k + \epsilon_{mjk}$$

where w_{jk} is the fetal weight and $\tilde{m}_{jk}$ is the latent variable for malformation for the j-th

fetus in litter k, and

$$\epsilon_{jk} = \begin{pmatrix} \epsilon_{wjk} \\ \epsilon_{mjk} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_w^2 & \tau\sigma_w\sigma_m \\ \tau\sigma_w\sigma_m & \sigma_m^2 \end{pmatrix}\right)$$

where σ_w^2 is the variance for w_{jk}, σ_m^2 is the variance for $\tilde{m}_{jk}$, and τ is the correlation between w_{jk} and $\tilde{m}_{jk}$ on the same fetus. A result of this formulation is that $\tilde{m}_{jk}|w_{jk}, d_k \sim N(\beta_0 + \beta_1 d_k + (\frac{\sigma_1}{\sigma_2})\tau(w_{jk} - (\alpha_0 + \alpha_1 d_k)), \sigma_m^2(1-\tau^2))$ which leads to the result:

$$P(m_{jk} = 1|w_{jk}, d_k) = \Phi\left(\frac{\beta_0 + \beta_1 d_k + (\frac{\sigma_1}{\sigma_2})\tau(w_{jk} - (\alpha_0 + \alpha_1 d_k))}{\sqrt{\sigma_m^2(1-\tau^2)}}\right)$$

In order for all coefficient parameters to be estimable, the model must be reparameterized to

$$P(m_{jk} = 1|m_{jk}, d_k) = \Phi(\beta_0^* + \beta_1^* d_k + \beta_2^*(w_{jk} - (\alpha_0 + \alpha_1 d_k)).$$

Of course, for the model to have practical use, it must be extended to also address intra-litter correlation. Fortunately, the multivariate normal makes including this correlation structure relatively easy to do. In Catalano-Ryan's extended model, the latent variable follows a multivariate normal with the following moments:

$$E\begin{pmatrix} \mathbf{w_k} \\ \tilde{\mathbf{m}}_\mathbf{k} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{d_k 1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{d_k 1} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \beta_0 \\ \beta_1 \end{pmatrix}$$

and

$$Var\begin{pmatrix} \mathbf{w_k} \\ \tilde{\mathbf{m}}_\mathbf{k} \end{pmatrix} = \begin{pmatrix} \sigma_w^2[(1-\rho_w)\mathbf{I} + \rho_w\mathbf{J}] & \sigma_w\sigma_m[(\tau - \rho_{wm}\mathbf{I} + \rho_{wm}\mathbf{J}] \\ \sigma_w\sigma_m[(\tau - \rho_{wm}\mathbf{I} + \rho_{wm}\mathbf{J}] & \sigma_m^2[(1-\rho_m)\mathbf{I} + \rho_m\mathbf{J}] \end{pmatrix}$$

where ρ_{wm} is the correlation between w_{jk} and $\tilde{m}_{jk}$, ρ_w is the intra-litter correlation for fetal weight, and ρ_w is the intra-litter correlation for the latent variables for malformation. Thus, the conditional distribution for malformations is $\tilde{\mathbf{m}}_k|\mathbf{w_k} \sim N(\boldsymbol{\mu}_{wk}, \sigma_m^2\Sigma_k)$

where

$$\mu_{wjk} = \beta_0 + \beta_1 d_k + \frac{\sigma_m}{\sigma_w}(\frac{\tau + (N_k - 1)\rho_{wm}}{1 + (N_k - 1)\rho_w})\bar{e}_{wk} + \frac{\sigma_m}{\sigma_w}(\frac{\tau - \rho_{wm}}{1 - \rho_w})(e_{wjk} - \bar{e}_{wk})$$
$$e_{wjk} = m_{jk} - (\alpha_0 + \alpha_1 d_k)$$
$$\bar{e}_{wk} = \bar{m}_k - (\alpha_0 + \alpha_1 d_k)$$
$$\Sigma_k = [(1 - \rho_w) - \frac{(\tau - \rho_{wm})^2}{1 - \rho_w}]\mathbf{I} + [\rho_m - \frac{(1 - \rho_m)(\tau^2 + (N_k - 1)\rho_{wm}^2) - (\tau - \rho_{mw})^2}{(1 - \rho_w)(1 + (N_k - 1)\rho_w)}]\mathbf{J}$$

After reparameterizing the model to ensure estimability, the model becomes $E[m_{jk}|\mathbf{w_k}] = \Phi(\beta_0^* + \beta_1^* d_k + \beta_2^* \bar{e}_{wk} + \beta_3^*(e_{wjk} - \bar{e}_{wk}))$. Note that e_{wjk} is the fetus-specific residual of the weight model and $\bar{e}_{wk}$ is the average of the litter weight residuals. Thus, according to this model, both individual fetal weight and average litter weight affect the probability of malformation. More specifically, when the average litter weight is lower than expected ($\hat{e}_{wjk}$ is negative), a fetus of that litter is more likely to be malformed. Similarly, when the weight of a fetus is lower than the average weight for that litter ($e_{wjk} - \hat{w}_{jk}$ is negative), the fetus will also have a higher malformation rate. Thus, these two additional parameters reflect the inherent correlation between malformation rate and fetal weight within a fetus, as well as intra-litter correlation.

Catalano and Ryan propose using two sets of estimating equations to estimate parameters:

$$\sum_{k=1}^{K} \mathbf{X_k^T V_{wk}^{-1}}(\mathbf{w_k} - \mathbf{X_k}\alpha) = 0$$
$$\sum_{k=1}^{K} \frac{\partial \mathbf{E}[\mathbf{m_k}|\mathbf{w_k}]}{\partial \boldsymbol{\beta}} \mathbf{V_{mk}^{-1}}(\mathbf{m_k} - \mathbf{E}[\mathbf{m_k}|\mathbf{w_k}]) = 0$$

Note that $\hat{\alpha}$ obtained from solving the first estimating equation, substitutes α in $E[\mathbf{m_k}|\mathbf{w_k}]$ for the second estimating equation without iteration.

While the model's latent formulation has an intuitive appeal, it also has some disadvantages. For instance, the regression parameters for the malformation model are conditional on fetal weight. Thus, the β parameters in this model do not have a marginal interpretation because of the nonlinear link function. Without a model that characterizes

marginal risk for malformation, it is not possible to calculate a univariate BMD for that response.

Fitzmaurice and Laird (Fitzmaurice and Laird, 1995) propose a similar model that reverses the role of weight and malformation conditioning in terms of characterizing the joint density. In the Fitzmaurice-Laird model, m_j follows a Bernoulli distribution and w_j follows a normal distribution, conditional on malformation status. That is, $m_j \sim Bernoulli(p_j)$ and $w_j|m_j \sim N(\mathbf{X}_j\boldsymbol{\alpha} + \gamma(m_j - p_j), \sigma^2)$, where $logit(p_j) = \mathbf{X_j}\boldsymbol{\beta}$ and γ is the parameter from a regression of w_j on m_j. Thus, in this model, the joint distribution is characterized as $f_{m_j,w_j}(m_j, w_j) = f_{w_j|m_j}(w_j|m_j)f_{m_j}(m)$, defined by the marginal distribution of malformation status and conditional distribution of fetal weight, whereas the Catalano-Ryan model defines the joint density as the product of the marginal distribution of fetal weight and conditional distribution of malformation status. The advantage of the Fitzmaurice-Laird model is, since $E[w_j|m_j] = \mathbf{X}_j\boldsymbol{\alpha}+\gamma(m_j-p_j), E[w_j] = E[E[w_j|m_j]] = \mathbf{X}_j\boldsymbol{\alpha}$. Thus, both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters have a marginal interpretation in this setting.

In the clustered setting, the mean models are as follows:

$$logit(\mathbf{p_k}) = \mathbf{X_k}\boldsymbol{\beta}$$
$$E[w_{jk}|\mathbf{m}_k] = \mathbf{X}_{jk}\boldsymbol{\alpha} + \gamma_1(m_{jk} - p_{jk}) + \gamma_2\sum_{j=1}^{n_k}(m_{jk} - p_{jk})$$

and the following GEEs are used to solve for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_2 = (\boldsymbol{\alpha}, \gamma_1, \gamma_2)$:

$$\sum_{k=1}^{K}\begin{pmatrix} \frac{\partial E[\mathbf{m}_k]}{\partial \boldsymbol{\beta}} & \frac{\partial E[\mathbf{m}_k]}{\partial \boldsymbol{\alpha}_2} \\ \frac{\partial E[\mathbf{w}_k|\mathbf{m}_k]}{\partial \boldsymbol{\beta}} & \frac{\partial E[\mathbf{w}_k|\mathbf{m}_k]}{\partial \boldsymbol{\alpha}_2} \end{pmatrix}^T Var^{-1}\begin{pmatrix} \mathbf{m}_k \\ \mathbf{w}_k|\mathbf{m}_k \end{pmatrix}\begin{pmatrix} \mathbf{m_k} - E[\mathbf{m}_k] \\ \mathbf{w_k} - E[\mathbf{w}_k|\mathbf{m}_k] \end{pmatrix} = 0$$

which can be expressed as:

$$\sum_{k=1}^{K}\begin{pmatrix} \mathbf{X}^T p_{m_k}(1-p_{m_k})\mathbf{I}_{n_k} & -(\gamma_1+\gamma_2)\mathbf{X}^T p_{m_k}(1-p_{m_k})\mathbf{I}_{n_k} \\ \mathbf{0} & \mathbf{W}^T \end{pmatrix}$$
$$\begin{pmatrix} \mathbf{V}_{m_k}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{w_k|m_k}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{m_k} - \mathbf{p_k} \\ \mathbf{w_k} - E[\mathbf{w}_k|\mathbf{m}_k] \end{pmatrix} = 0$$

where $\mathbf{W}_k = (\mathbf{W}_{1k}, \mathbf{W}_{2k}, ..., \mathbf{W}_{jk}, ..., \mathbf{W}_{n_k k})$, and $\mathbf{W}_{jk} = (\mathbf{X}_{jk}, m_{jk} - p_{jk}, \sum_{j=1}^{n_k}(m_{jk} - p_{jk}))$ (Fitzmaurice and Laird, 1995). Note that, while the malformation model only de-

pends on β, the weight model is conditional on p_{m_k} and therefore on β as well as α_2, which is why $\frac{\partial E[\mathbf{m}_k]}{\partial \boldsymbol{\alpha}_2}$ is $\mathbf{0}$ while $\frac{\partial E[\mathbf{w}_k|\mathbf{m_k}]}{\partial \boldsymbol{\alpha}_2}$ is not, leading to a non-symmetric matrix for D^T.

In order to calculate a joint BMD, it is necessary to characterize joint risk through a likelihood and to estimate the parameters that characterize the correlation between the outcomes. The Fitzmaurice-Laird model only defines likelihoods for the malformation and weight conditional on malformation. Furthermore, the model accounts for correlation between fetal weight and malformation through the parameters γ_1 and γ_2 in the conditional weight model but these parameters do not directly estimate the correlation between weight and malformation. On the other hand, the Catalano-Ryan model does specify the joint probability, but the correlation parameter, τ, which describes the relationship between fetal weight and malformation is not directly estimated. Thus, neither method can be used calculate a joint BMD.

Regan and Catalano (Regan and Catalano, 1999) propose a method which retains estimation of the inter-outcome correlation but also takes advantage of the robust properties of GEE. Like the Catalano-Ryan model, the distribution of w and m are determined by the distribution of w and latent variable $\tilde{m}$ which follow a bivariate normal distribution. Without loss of generality, $\tilde{m}_{jk}$ is further standardized so that σ_m = 1. Thus, the density function for the weight and latent malformation variable is

$$f(w_{jk}, \tilde{m}_{jk}) = \frac{1}{2\pi\sigma_w\sqrt{1-\tau^2}} \times \\ exp\left(\frac{-1}{2(1-\tau^2)}\left[\left(\frac{w_{jk}-\mu_w}{\sigma_w}\right)^2 - 2\tau\left(\frac{w_{jk}-\mu_w}{\sigma_w}\right)(\tilde{m}_{jk}-\gamma_m) + (\tilde{m}_{jk}-\gamma_m)^2\right]\right)$$

where $\gamma_m = \mu_m/\sigma_w$, the mean of the standardized $\tilde{m_{jk}}$. From this density, it can be shown that $f(w_{jk}, m_{jk}) = \Phi(\gamma_{m|w_{jk}})^{m_{jk}}[1 - \Phi(\gamma_{m|w_{jk}})^{(1-m_{jk})}]f(w_{jk})$ where $\gamma_{m|w_{jk}} = \frac{\gamma_m + \tau\frac{w_{jk}-\mu_w}{\sigma_w}}{\sqrt{1-\tau^2}}$. Also note that, in addition to mean weight and malformation, weight variance and inter-outcome correlation can also be modeled as dose-dependent. Previously proposed models have assumed these parameters to be non-dose-dependent. However, in developmental toxicity data, it is often the case that the negative correlation between malformation and fetal weight grows stronger, and that fetal weight variance increases, as dose in-

creases (Chen and Gaylor, 1992).

Therefore, the dose response components in this model are as follows:

$$\begin{aligned} \gamma_{m_{jk}} &= \mathbf{X}_{jk}^T\boldsymbol{\beta} \qquad \tau_{jk} = \frac{e^{\mathbf{X}_{jk}^T\boldsymbol{\theta}}-1}{e^{\mathbf{X}_{jk}^T\boldsymbol{\theta}}+1} \\ \mu_{w_{jk}} &= \mathbf{X}_{jk}^T\boldsymbol{\alpha} \quad log(\sigma^2_{w_{jk}}) = \mathbf{X}_{jk}^T\boldsymbol{\eta} \end{aligned}$$

To account for clustering, the parameters are estimated using the the following GEE which uses a working covariance matrix to allow for the possible misspecification of correlation:

$$\sum_{k=1}^{K} \begin{pmatrix} \frac{\partial E[\mathbf{m_k}|\mathbf{w_k}]}{\partial\beta} & 0 & 0 \\ \frac{\partial E[\mathbf{m_k}|\mathbf{w_k}]}{\partial\theta} & 0 & 0 \\ \frac{\partial E[\mathbf{m_k}|\mathbf{w_k}]}{\partial\alpha} & \frac{E[\mathbf{w_k}]}{\partial\alpha} & 0 \\ \frac{\partial E[\mathbf{m_k}|\mathbf{w_k}]}{\partial\eta} & 0 & \frac{E[\mathbf{s_k}]}{\partial\eta} \end{pmatrix} \begin{pmatrix} \mathbf{V}_{m_k} & \mathbf{V}_{wm_k} & \mathbf{0} \\ \mathbf{V}_{wm_k} & \mathbf{V}_{w_k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}_{s_k} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{m_k} - \Phi(\boldsymbol{\gamma}_{m|w_k}) \\ \mathbf{w_k} - \boldsymbol{\mu}_{w_k} \\ \mathbf{s_k} - \boldsymbol{\sigma}^2_{w_k} \end{pmatrix} = 0$$

where

$$\begin{aligned} s_{jk} &= (w_{jk} - \mu_w)^2 \\ \mathbf{V}_{w_k} &= \Sigma_{w_k}^{1/2}[(1-\rho_w)\mathbf{I}_{\mathbf{n_j}} + \rho_w\mathbf{J}_{\mathbf{n_k}}]\Sigma_{w_k}^{1/2} \\ \mathbf{V}_{m_k} &= \Sigma_{m_k}^{1/2}[(1-\rho_m)\mathbf{I}_{\mathbf{m_j}} + \rho_m\mathbf{J}_{\mathbf{n_k}}]\Sigma_{m_k}^{1/2} \\ \mathbf{V}_{S_k} &= \Sigma_{s_k}^{1/2}[(1-\rho_m)\mathbf{I}_{\mathbf{m_j}} + \rho_m\mathbf{J}_{\mathbf{n_k}}]\Sigma_{s_k}^{1/2}/\phi_m \\ \mathbf{V}_{wm_k} &= \Sigma_{w_k}^{1/2}[-\rho_{wm}\mathbf{I}_{\mathbf{n_k}} + \rho_{wm}\mathbf{J}_{\mathbf{n_k}}]\Sigma_{m_k}^{1/2}/\phi_m^{1/2} \end{aligned}$$

and

$$\begin{aligned} \Sigma_{m_k} &= diag[\Phi(\gamma_{m|w_{jk}})(1-\Phi(\gamma_{m|w_{jk}}))] \\ \Sigma_{w_k} &= diag[\sigma^2_{w_{jk}}] \\ \Sigma_{s_k} &= diag[2\sigma^4_{w_{jk}}] \end{aligned}$$

and ϕ_m is a scale parameter. Method of moments estimation is used to calculate estimates for the correlation parameters τ_w, ρ_m and ρ_{wm}, as well as ϕ_m.

Regan and Catalano use the assumed joint likelihood model for risk assessment. Because the inter-outcome correlation is also estimated, a joint BMD can be calculated. Let

us assume that a malformation or a fetal weight lower than the cutpoint w_c is considered an adverse event. Then,

$$p(dose) = P(\text{adverse event at dose}) = 1 - P((m_{jk} = 1) \cap (w_{jk} > w_c)) =$$
$$1 - \int_{-\infty}^{\gamma_m(d)} \int_{w_c}^{\infty} \phi_2(w_{jk}, m_{jk} | \mu_w = \mu_w(dose), \mu_{\tilde{m}} = 0, \sigma_w = \sigma_w, \sigma_{\tilde{m}} = 1, \tau = \tau) dw_{jk} d\tilde{m}_{jk}$$

where ϕ_2 is the bivariate normal density function and μ_w, σ_w, and τ are functions of dose. This formula reduces to

$$\Phi(\gamma_m(dose)) + \Phi_2\left(-\gamma_m(dose), \frac{w_c - \mu_w(dose)}{\sigma_w(dose)} | \tau(dose)\right)$$

where Φ and Φ_2 are the cumulative distribution functions for the standard univariate and standard bivariate normal distribution, respectively. This joint risk formulation can be used to calculate a joint BMD and associated BMDL.

Molenberghs, Geys, and Buyse (Molenberghs et al., 2001) propose an entirely different model, based on the Plackett-Dale distribution rather than a latent bivariate normal distribution. Let $F_{w_k}(x) =$ cumulative distribution function for w_k and let $F_{m_k}(y) =$ cumulative distribution function for m_k . Under the Plackett-Dale model, the joint cumulative distribution function for m_k and w_k is

$$F_{w_k,m_k} = \begin{cases} \frac{1+(F_{w_k}+F_{m_k})(\psi_k-1)-S(F_{w_k},F_{m_k},\psi_k)}{2(\psi_k-1)} & \psi_k \neq 1 \\ F_{w_k}F_{m_k} & \psi_k = 1 \end{cases}$$

where

$$S(F_{w_k}, F_{m_k}, \psi_k) = \sqrt{[1 + (\psi_k - 1)(F_{w_k} + F_{m_k})]^2 + 4\psi_k(1 - \psi_k)F_{w_k}F_{m_k}}$$

ψ_k, known as the global cross-ratio, defines the dependence structure of w_k and m_k,

$$\psi_k = \frac{F_{w_k,m_k}(1 - F_{w_k} - F_{m_k} + F_{w_k,m_k})}{(F_{w_k} - F_{w_k,m_k})(F_{m_k} - F_{w_k,m_k})}$$

and is used to derive the above joint cumulative density function.

From this definition, the joint density function is derived to be

$$f_{w_k,m_k}(w,0) = \begin{cases} \frac{f_{w_k}(w)}{2}[1 - \frac{1+F_{w_k}(w)(\psi_k-1)-F_{m_k}(0)(\psi_k+1)}{S(F_{w_k}(w),F_{m_k}(0),\psi_k}] & \psi_k \neq 1 \\ f_{w_k}(w)(1-p_k) & \psi_k = 1 \end{cases}$$
$$f_{w_k,m_k}(w,1) = f_{w_k}(w) - f_{w_k,m_k}(w,0)$$

To account for clustering, a pseudo-likelihood score function, $pl = \sum_{k=1}^{K}\sum_{j=1}^{J} ln(f_{w_{jk},m_{jk}}(w,m))$, rather than a full likelihood score function, is used for computational simplicity and stability.

Letting $\theta_{\mathbf{jk}} = (\mu_{jk}, \sigma^2_{jk}, \pi_{jk}, \psi_{jk})^T$, the vector of parameters of interest, we can characterize the dose response model as

$$\eta_{jk} = \begin{pmatrix} \mu_{jk} \\ ln(\sigma^2_{jk}) \\ logit(p_{jk}) \\ ln(\psi_{jk}) \end{pmatrix} = \mathbf{X}_{jk}\beta$$

Estimates for β can be obtained from solving the following score-based estimating function:

$$\mathbf{U}(\beta) = \sum_{k=1}^{K} \mathbf{U_k}(\beta) = \sum_{k=1}^{K}\sum_{j=1}^{n_k} \left(\frac{\partial \eta_k}{\partial \beta}\right)^T \left(\frac{\partial \eta_k}{\partial \theta_k}\right)^{-T} \left(\frac{\partial ln(f_{w_{jk},m_{jk}}(x,y))}{\partial \theta_i}\right) = 0$$

The dependence structure for the two outcomes is defined by the global cross-cut ratio, ψ_k. Thus, this approach allows for great flexibility in choice of marginal distributions. The global cross-ratio can be interpreted as the odds-ratio comparing malformation odds and fetal weight odds, where fetal weight is thought of as a dichotomized variable using an unestimated cut point, w_c. That is, when $\psi_k = 1$, the two outcome are independent, when $\psi_k > 0$, there is a positive correlation between weight and malformation, and when $\psi_k < 0$, there is a negative correlation. Thus, the model's characterization of the association is completely different than that of the probit model, which uses a correlation parameter from a multivariate normal distribution.

Again, assuming observing a malformation or a fetal weight lower than the cutpoint w_c is considered an adverse event, the probability of an adverse event can be characterized as

$$\begin{aligned} P(w < w_c \cup m = 1|dose) &= p(m = 1|dose) + p(w < w_c|dose) - p(m = 1 \cap w < w_c|dose) \\ &= p(m = 1|dose) + F_{w,m}(w_c, 0|\psi(dose)). \end{aligned}$$

Using this formulation of the joint risk, a joint BMD can be calculated (Geys et al., 2001).

1.6 Hierarchical Outcomes

The models described above have mostly focused on outcomes from live fetuses, namely, fetal weight and malformation or multiple types of malformations. Another important outcome that toxicologists consider is early prenatal loss and fetal death, often collectively termed embryolethality. The addition of death as an outcome of interest presents a new statistical challenge because of the hierarchical relationship between death and live outcomes in the litter. In particular, being able to observe malformation status and fetal weight is conditional on the fetus being alive at the time of sacrifice. Say we are only interested in deaths and malformations as outcomes. Within a dam, one can easily estimate $p(d)$ with D_k/n_k and, similarly, $p(m|\bar{d})$ with M_k/l_k, where D_k and M_k are the number of deaths and malformations observed respectively, n_k is the number of implants in dam k and $l_k = n_k - D_k$ is the number of live fetuses (litter size) for dam k. However, it is more difficult to characterize the joint risk of both death and malformation. Similarly, any risk statement we can make on fetal malformation is conditional on those fetuses surviving gestation. Many methods obviate this road block by assuming conditional independence: that d and $m, w|d$ are independent. This assumption simplifies the construction of joint risk from multiple outcomes to $P(d)P(m, w|\bar{d})$. The calculation may be appropriate for univariate unclustered hierarchical outcomes because observing a death in one animal would not inform the malformation rate or fetal weight for a different animal. However, this logic breaks down in litter data where death rate of a litter is expected to inform the malformation status and fetal weights of animals in the same litter.

For an example of a hierarchical model that assumes conditional independence, Catalano, Ryan and Scharfstein (Catalano et al., 1994) present a method to model the dose response for two hierarchical binary outcomes. In their approach, two dose-response models are fitted: one for death and one for malformation, with the malformation model being conditional the fetus surviving the gestation period. Letting $p_d(dose)$ be the probability of death at the specified dose and $p_m(dose)$ be the probability of malformation, two

sets of GEEs are solved simultaneously:

$$\sum_{k=1}^{K}\sum_{j=1}^{n_k}\frac{\partial p_d(dose)}{\partial \beta_d}\mathbf{V}_{d_{jk}}^{-1}(\mathbf{d}_{jk}-p_d(dose)\mathbf{1}_{n_k})=0$$
$$\sum_{k=1}^{K}\sum_{j=1}^{n_k-D_k}\frac{\partial p_m(dose)}{\partial \beta_m}\mathbf{V}_{m_{jk}}^{-1}(\mathbf{m}_{jk}-p_m(dose)\mathbf{1}_{n_k-D_k})=0$$

where $\mathbf{V}_{d_{jk}}$ and $\mathbf{V}_{m_{jk}}$ are the assumed covariance matrices for death and malformations, respectively, β_d and β_m are the parameters from the death and malformation dose-response model, and $\mathbf{d}_{jk}$ is a length n_k vector indicating death for fetuses from dam jk while $\mathbf{m}_{jk}$ is a length $n_k - D_k$ vector indicating malformations for live fetuses from dam jk.

In order to perform a risk assessment analysis, they define an adverse event as either observing a death or malformation. Assuming conditional independence, $P(d \cup m) = 1 - P(\bar{d} \cap \bar{m}) = 1 - P(\bar{d})P(\bar{m}|\bar{d}) = 1 - (1 - P(d))(1 - P(m|\bar{d}))$. Thus, since they assume d and $m|\bar{d}$ are fully independent, they can fit a model with death as the outcome of interest where number of implants is considered the denominator, while also fitting a separate model treating malformations as the outcome of interest where number living fetuses is considered the denominator, to calculate the overall probability of an adverse event. This probability can then be used to calculate a BMD. Catalano, Scharfstein, *et al.* extend this approach to include fetal weight as an outcome of interest by modeling the live outcomes (malformations and weight) together using the Catalano-Ryan probit model (Catalano et al., 1993).

Another example of a model that incorporates lethality as an outcome is the Dirichlet-trinomial model (Chen et al., 1991), which extends the beta-binomial model to include two binary outcomes, malformation and death, by replacing the binomial with a trinomial and beta with a Dirichlet distribution. The resulting, more general, hierarchical model can be also used for calculating a joint BMD. Note that the assumption of conditional independence is implicit in the likelihood formulation. However, both of these papers ignore the inter-outcome litter effects that may be present in the hierarchical outcomes.

In situations without clustered data, conditional independence is a reasonable assumption. One would expect that outcomes from one animal would not inform the likelihood for a live fetus from another animal. However, when fetuses are clustered into litters, as in developmental or other animal studies involving litter data, it is possible that knowing the death experience of a given litter can affect the malformation rate and fetal weight distribution of the remaining live outcomes in the same litter. In particular, one might expect that a litter with a high death rate could also have a high malformation rate and a lower fetal weight distribution for the live fetuses. In practice, and somewhat contradictory to the conditional independence assumption that many models employ, litter size is often included as a covariate in regression models for live-outcomes as an ad-hoc method for taking into account the possibility that death rate may affect malformation rates and fetal weights (Chen, 1993).

Christensen (Christensen, 2004) formalizes this ad hoc approach and extends the Ochi-Prentice model by including fetal death as a variable in such a way that it does not assume conditional independence. Essentially, three possible outcomes, no adverse event, malformation, and death are treated as ordinal events. In Christensen's model, two threshold parameters, τ_m and τ_d, are used to define how the latent variable relates to the observed outcomes. Letting $\tilde{y}_{jk}$ be the latent variable for fetus j from dam k, if $\tilde{y}_{jk} < \tau_m$, then no adverse event is observed for that fetus, if $\tau_m < \tilde{y}_{jk} < \tau_d$, then a malformation is observed for the fetus and if $\tilde{y}_{jk} > \tau_d$, a fetal death is observed. Like the Ochi-Prentice model, $\tilde{\mathbf{y}}_k$, the vector denoting the latent variables for the fetuses from dam k, follows a multivariate normal distribution with mean $\mu\mathbf{1}_n$ and variance $\sigma^2((1-\rho)\mathbf{I}_n + \rho\mathbf{J}_n)$. Letting H_k denote the number of healthy fetuses from litter k, the joint distribution of the three outcomes from litter k can be expressed as:

$$P(H_k, M_k, D_k) \propto \int_B \phi_n(\tilde{\mathbf{z}}_k|0, 1, \rho)d\tilde{\mathbf{z}}_k$$

where

$$\begin{aligned}
\tilde{\mathbf{z}}_k &= \tilde{\mathbf{y}}_k - \mathbf{1}_{n_k}\mu \\
B &= (\tilde{\mathbf{z}}_k | (\tilde{z}_k < -\gamma_m, k \leq H_k) \cup (\gamma_m \leq \tilde{z}_{jk} < -\gamma_d, H_k < k < H_k + M_k) \\
&\quad \cup (\tilde{z}_{jk} \geq -\gamma_d, j > H_k + M_k))) \\
\gamma_m &= \tau_m - \mu \\
\gamma_d &= \tau_d - \mu
\end{aligned}$$

γ_m and γ_d are standardized cutpoints.

From this formulation, it follows that the probablity of death is $\Phi(\gamma_d)$ and probability of malformation is $\Phi(\gamma_m) - \Phi(\gamma_d)$. Using the above likelihood, the model specification is as follows:

$$\begin{aligned}
\mu &= f(\mathbf{X}_1\boldsymbol{\beta}) \\
\gamma_m &= -\tau_m + \beta_0 + f(\mathbf{X_1}\boldsymbol{\beta}) = \tau_m^*(\mathbf{X}_m\boldsymbol{\lambda}_m) + f(\mathbf{X}_1\boldsymbol{\beta}) \\
\gamma_D &= -\tau_D + \beta_0 + f(\mathbf{X_1}\boldsymbol{\beta}) = \tau_d^*(\mathbf{X}_d\boldsymbol{\lambda}_d) + f(\mathbf{X}_1\boldsymbol{\beta}) \\
\rho &= g(\mathbf{X}_2\boldsymbol{\xi})
\end{aligned}$$

where $g(\cdot)$ can either be the identity function or Fisher's Z-transformation.

In order to compute MLE's for the parameters of interest, it is necessary to take derivatives of the log-likelihood. The derivatives of the log-likelihood with respect to μ and ρ can be quite complicated as they involve differentiating the following integrals:

$$\begin{aligned}
&\int_B \sum_{j=1}^{n_k} \tilde{z}_{jk}\phi_n(\tilde{\mathbf{z}}_k|0,1,\rho)d\tilde{\mathbf{z}}_k \\
&\int_B \sum_{j=1}^{n_k}\sum_{j'=1}^{n_k} \tilde{z}_{jk}\tilde{z}_{j'k}\phi_n(\tilde{\mathbf{z}}_k|0,1,\rho)d\tilde{\mathbf{z}}_k.
\end{aligned}$$

As in the Ochi-Prentice model, moment generating functions are used to express the integrals as the product of univariate normal density function and a (n-1)-dimensional

multivariate normal integral, conditional on one of the two thresholds. Using this method, we can calculate the first and second derivatives of interest for the likelihood.

As with the Ochi-Prentice model, approximations developed by Mendell and Elston (Mendell and Elston, 1974) are used to carry out the calculations. Using this method, the joint risk of the outcomes of a litter:

$$P(\tilde{z}_1 < -\gamma_m, ..., \tilde{z}_{M_k} < -\gamma_m, \tilde{z}_{M_k+1} < -\gamma_d, ..., \tilde{z}_{M_k+D_k} < -\gamma_d) =$$
$$\Pi_{k=1}^{M_k} P(\tilde{z}_k < -\gamma_m | \tilde{z}_k < -\gamma_m, ..., \tilde{z}_{k-1} < -\gamma_m) \times$$
$$\Pi_{j=M_k+1}^{M_k+D_k} P(\tilde{z}_j < -\gamma_d | \tilde{z} < -\gamma_m, ..., \tilde{z}_{M_k} < -\gamma, \tilde{z}_{M_k+1} < -\gamma_d, ..., \tilde{z}_{j-1} < -\gamma_d)$$

can be approximated by $\Pi_{j=1}^{M_j} \Phi(q_j) \Pi_{k=M_j+1}^{M_j+D_j} \Phi(u_j)$.
For the first product, $w_j = 1, w_j + 1 = (w_j - a_{wj} r_j)/\sigma_j, r_j = (r_{j-2} - 1)/\sigma_{j-1}^2 + 1, w_j = -\phi(w_j)/\Phi(w_j)$ and $\sigma_j^2 = 1 - r_j^2 a_{wj}(a_{wj} - w_j)$ for $j = 1, ..., M_k$. For the second product, u_j must be calculated in two stages. First, for $u_1, ..., u_{M_k+1}$, $u_1 = -\gamma_d$ and $u_{s+1} = u_t - w_{ws} r_t/\sigma_s$, where a_{ws}, r_s, and σ_s^2 are the same as above. For $u_{M_j+1}, ..., u_{M_j+D_j}$, $u_s + 1 = (u_s - a_{us} r_s)/\sigma_s$ where $r_s = (r_{s-1} - 1)/\sigma_{s-1}^2 + 1$, $a_{us} = -\phi(u_s)/\Phi(u_s)$, and $\sigma = 1 - r_s^2 a_{us}(a_{us} - u_s)$.

Christensen also extends this model to include fetal weight by assuming fetal weight and the latent variable follow a multivariate normal distribution. In this case, regression models can be specified for all parameters of interest: fetal weight mean and variance, both litter-effect correlations, and the inter-outcome correlation.

The likelihood for this model can be extremely complicated, and even using approximation techniques to make certain calculations more tractable, the method can be computationally intensive, sensitive to starting values, and unstable with certain outcome data patterns. Also, as a likelihood method, it is not robust to model misspecification. However, it is one of the first models that allows joint risk assessment without assuming conditional independence.

As mentioned above, it is common practice to adjust the live-outcome models by the litter's death rate or litter size to informally take into account the conditional independence in the observed data. While Christensen's model formally incorporates conditional

dependence into his model, its complexity makes it unappealing for many researchers. Carey (Carey, 2006) develops a model which relaxes the conditional independence assumption but also retains the simplicity and accessibility of the more commonly used ad-hoc methods. Essentially, Carey uses adjustment covariates for conditional models that are derived from a proper likelihood model that does not assume conditional independence.

Unlike Christensen's method, Carey's likelihood uses two latent variables, one for death and one for malformation, denoted as $\tilde{d}$ and $\tilde{m}$ respectively. The two latent variables and fetal weight follow a multivariate normal distribution. More specifically, for the k-th litter:

$$\begin{pmatrix} \tilde{\mathbf{d}}_k \\ \mathbf{w}_k \\ \tilde{\mathbf{m}}_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_w \\ \mu_m \end{pmatrix}, \begin{pmatrix} \Sigma_d & \Sigma_{dw} & \Sigma_{dm} \\ \Sigma_{dw} & \Sigma_w & \Sigma_{wm} \\ \Sigma_{dm} & \Sigma_{wm} & \Sigma_m \end{pmatrix} \right)$$

where

$$\begin{aligned}
\mu_d &= (\tilde{\alpha}_0 + \tilde{\alpha}_1 dose_k)\mathbf{1}_{n_k} \\
\mu_w &= (\beta_0 + \beta_1 dose_k)\mathbf{1}_{l_k} \\
\mu_m &= (\tilde{\eta}_0 + \tilde{\eta}_1 dose_k)\mathbf{1}_{l_k} \\
\Sigma_d &= \sigma_d^2((1-\rho_d)\mathbf{I}_{n_k} + \rho_d \mathbf{J}_{n_k}) \\
\Sigma_w &= \sigma_w^2((1-\rho_w)\mathbf{I}_{l_k} + \rho_w \mathbf{J}_{l_k}) \\
\Sigma_m &= \sigma_m^2((1-\rho_m)\mathbf{I}_{l_k} + \rho_m \mathbf{J}_{l_k}) \\
\Sigma_{dw} &= \Sigma_{wd}^T = \rho_{wd}\sigma_w\sigma_d \mathbf{J}_{n_k \times l_k} \\
\Sigma_{dm} &= \Sigma_{md}^T = \rho_{md}\sigma_m\sigma_d \mathbf{J}_{n_k \times l_k} \\
\Sigma_{wm} &= \Sigma_{mw}^T = \rho_{wm}\sigma_w\sigma_m \mathbf{J}_{l_k}
\end{aligned}$$

and l_k denotes the number of live fetuses while n_k denotes the number of implants in litter k.

As weight is observed only when death does not occur for a fetus, it may be of greater interest to consider $\mathbf{w}|\mathbf{d}$. Similarly, the conditional distribution $\mathbf{m}|\mathbf{d}, \mathbf{w}$ may be

more attractive than the marginal distribution of malformation. Unlike the marginal distributions, the conditional distributions $\mathbf{w}|\mathbf{d}$ and $\mathbf{m}|\mathbf{w},\mathbf{d}$ both take into account the death outcomes of the litter and therefore may inform how death from the litter relates to malformation rate and fetal weight.

Given the above likelihood, the marginal distribution of death and conditional distribution of fetal weight and malformation can be expressed as:

$$\begin{pmatrix} \tilde{\mathbf{d}}_k \\ \mathbf{w}_k|\tilde{\mathbf{d}}_k \\ \tilde{\mathbf{m}}_k|\mathbf{w}_k,\tilde{\mathbf{d}}_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_{w|d} \\ \mu_{m|w,d} \end{pmatrix}, \begin{pmatrix} \Sigma_d & \mathbf{0}_{n_k\times l_k} & \mathbf{0}_{n_k\times l_k} \\ \mathbf{0}_{l_k\times n_k} & \Sigma_{w|d} & \mathbf{0}_{l_k} \\ \mathbf{0}_{l_k\times n_k} & \mathbf{0}_{l_k} & \Sigma_{m|w,d} \end{pmatrix}\right)$$

Note that, after conditioning on the means, the conditional and marginal outcomes are assumed to be uncorrelated. For example, how $\tilde{\mathbf{d}}$ informs $\mathbf{w}|\tilde{\mathbf{d}}$ is taken into account via the conditional mean $\mu_{w|d}$ and conditional variance $\Sigma_{w|d}$.

Now, $\mu_{w|d}$ and $\mu_{m|w,d}$ can be expressed as

$$\begin{aligned}
\mu_{w|d} &= (\beta_0 + \beta_1 dose) \\
&\quad + (\rho_{wd}\sigma_w)(1+\rho_d(n_k-1))^{-1}\left(\frac{\Sigma_{j=1}^{n_k}\tilde{d}_{ij} - n_k(\tilde{\alpha}_0+\tilde{\alpha}_1)}{\sigma_d}\right) \\
\mu_{m|w,d} &= (\tilde{\eta}_0 + \tilde{\eta}_1 dose) \\
&\quad + \frac{\sigma_m[\rho_{md}(1+\rho_w(l_k-1)) - \rho_{mw}\rho_{wd}l_k]}{(1-\rho_w(l_k-1))(1+\rho_d(n_k-1)) - \rho_{wd}^2 l_k n_k}\left(\frac{\Sigma_{j=1}^{n_k}\tilde{d}_{jk} - n_k(\tilde{\alpha}_0 - \tilde{\alpha}_1 dose)}{\sigma_d}\right) \\
&\quad + \frac{\sigma_m[\rho_{mw}(1+\rho_d(n_k-1)) - \rho_{mw}\rho_{wd}n_k]}{(1+\rho_w(n_k-1))(1+\rho_w(l_k-1)) - \rho_{wd}^2 n_k l_k}\left(\frac{\Sigma_{j=1}^{l_k} w_{jk} - l_k(\beta_0 - \beta_1 dose)}{\sigma_w}\right)
\end{aligned}$$

Note that $\mu_{w|d}$ can be expressed as the sum of marginal model for weight plus an adjustment covariate. More specifically, this additional adjustment covariate is a function of the mean standardized residuals for fetal death. Similarly, $\mu_{m|w,d}$ can be expressed as the sum of the marginal model for malformations plus two additional adjustment covariates, one a function of the mean standardized deaths and the other a function of the mean standardized weights. Unfortunately, these adjustment terms are quite complicated and include parameters from the latent theory that are not estimable. However, these theoretical models can be used to motivate simpler adjustment terms.

First, the latent death model must be rewritten to reflect that death is observed as a binary outcome. Thus, $\mu_d = E[d_{jk}] = P(d_{jk} = 1) = P(\tilde{d}_{jk} > 0) = \Phi(\frac{\tilde{d}_{jk}}{\sigma_d} > 0) = \Phi(\frac{\alpha_0}{\sigma_d} + \frac{\alpha_1}{\sigma_d} dose) = \Phi(\alpha_0 + \alpha_1 dose)$ where the $\tilde{\alpha}_0$ and $\tilde{\alpha}_1$ are reparameterized to α_0 and α_1 to ensure estimability.

From here, the conditional weight model can be expressed as:

$$\mu_{w|d} = \beta_0 + \beta_1 dose + \beta_2 (1 + \rho_d (n_k - 1))^{-1} \left(\frac{\Sigma_{j=1}^{n_k} d_{jk}/n_{jk} - n_k(\alpha_0 + \alpha_1)}{\sqrt{\Phi(\tilde{\alpha}_0 - \tilde{\alpha}_1 dose)[1 - \Phi(\alpha_0 + \alpha_1 dose)]/n_k}} \right)$$

where $\rho_{wd}\sigma_w$ is taken to be a single parameter, β_2 and ρ_d is estimated using the method of moments from the residuals of the fitted dose-response model.

For the conditional malformation model the theoretical adjustment covariates are quite complicated and include parameters from the latent model that are not estimable. However, this theoretical mean model can be used to motivate simpler adjustment terms and helps justify models that previously used ad-hoc approaches. Specifically, Carey derives adjusted covariates based on a first order bivariate Taylor expansion around the mean number of implants and mean litter size. In addition, like the marginal death model, a reparameteriztion of the parameters is necessary for estimability. This approximation to the conditional mean is expressed as:

$$\begin{aligned}
\mu_{m|w,d} = (\eta_0 + \eta_1 dose) &+ \eta_2 \left(\frac{\bar{d}_k - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)}{\sqrt{\Phi(\hat{\alpha}_0 - \hat{\alpha}_1 dose)[1 - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)]/n_k}} \right) \\
&+ \eta_3 (D_k - \bar{D}) \left(\frac{\bar{d}_k - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)}{\sqrt{\Phi(\hat{\alpha}_0 - \hat{\alpha}_1 dose)[1 - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)]/n_k}} \right) \\
&+ \eta_4 (n_k - \bar{n}) \left(\frac{\bar{d}_k - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)}{\sqrt{\Phi(\hat{\alpha}_0 - \hat{\alpha}_1 dose)[1 - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)]/n_k}} \right) \\
&+ \eta_5 \left(\frac{\bar{w}_k - \hat{\mu}_w}{\hat{\sigma}_w/\sqrt{l_k}} \right) \\
&+ \eta_6 (n_k - \bar{n}) \left(\frac{\bar{w}_k - \hat{\mu}_w}{\hat{\sigma}_w/\sqrt{l_k}} \right) \\
&+ \eta_7 (D_k - \bar{D}) \left(\frac{\bar{w}_k - \hat{\mu}_w}{\hat{\sigma}_w/\sqrt{l_k}} \right)
\end{aligned}$$

The covariates for the parameters η_2 and η_5 characterize the main effect of death ex-

perience of the litter and fetal weight, respectively. The other four covariates can be interpreted as interaction effects in which the main effects are multiplied by either the litter's deviation from the average number of implants or the litter's deviation from the average number of deaths.

Given these marginal and conditional models, and using the following dose-response framework:

$$\begin{aligned} E[d_{jk}]/\sqrt{Var(d_{jk})} &= \Phi(\alpha_0 + \alpha_1 dose_k) \\ E[m_{jk}]/\sqrt{Var(m_{jk})} &= \Phi(\eta_0 + \eta_1 dose_k) \\ E[w_{jk}] &= \beta_0 + \beta_1 dose_k \end{aligned}$$

we can fit the following GEE:

$$\sum_{k=1}^{I} \begin{pmatrix} \frac{\partial E(\mathbf{d}_k)}{\partial \alpha} & \mathbf{0} & \mathbf{0} \\ \frac{\partial E(\mathbf{w}_k|\mathbf{d}_k)}{\partial \alpha} & \frac{\partial E(\mathbf{w}_k|\mathbf{d}_k)}{\partial \beta} & \mathbf{0} \\ \frac{\partial E(\mathbf{w}_k|\mathbf{w}_k,\mathbf{d}_k)}{\partial \alpha} & \frac{\partial E(\mathbf{w}_k|\mathbf{w}_k,\mathbf{d}_k)}{\partial \beta} & \mathbf{0} \end{pmatrix}^T \begin{pmatrix} \mathbf{V}_{d_k} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{w_k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}_{m_k} \end{pmatrix}^{-1} \times \begin{pmatrix} \mathbf{d}_k - \Phi(\alpha_0 + \alpha_1 dose_k)\mathbf{1}_{n_k} \\ \mathbf{w}_k - E(\mathbf{w}_k|\mathbf{d}_k) \\ \mathbf{m}_k - E(\mathbf{m}_k|\mathbf{w}_k, \mathbf{d}_k) \end{pmatrix} = 0$$

where

$$\begin{aligned} \mathbf{V}_{d_k} &= \Sigma_{d_k}^{1/2}[(1-\rho_d)\mathbf{I}_{n_k} + \rho_d \mathbf{J}_{\mathbf{n_k}}]\Sigma_{d_k}^{1/2}/\phi_d \\ \mathbf{V}_{w_k} &= \Sigma_{w_k}^{1/2}[(1-\rho_w)\mathbf{I}_{l_k} + \rho_w \mathbf{J}_{\mathbf{l_k}}]\Sigma_{l_k}^{1/2} \\ \mathbf{V}_{m_k} &= \Sigma_{m_k}^{1/2}[(1-\rho_m)\mathbf{I}_{l_k} + \rho_m \mathbf{J}_{\mathbf{l_k}}]\Sigma_{m_k}^{1/2}/\phi_m \end{aligned}$$

and

$$\begin{aligned} \Sigma_{d_k} &= diag[\Phi(\alpha_0 + \alpha_1 dose)(1 - \Phi(\alpha_0 + \alpha_1 dose))] \\ \Sigma_{w_k} &= diag[\sigma^2_{w|d}] \\ \Sigma_{m_k} &= diag[\Phi(\eta_0 + \eta_1 dose)(1 - \Phi(\eta_0 + \eta_1 dose))] \end{aligned}$$

Depending on the specific data set, not all adjustment covariates for the conditional malformation model may be significant, especially given the high potential for collinearity among them. The interaction terms, in general, do not tend to be significant, but both main effect adjustment terms tend to be informative. What covariates one decides to include in the mean model should ultimately depend on the data. While using adjustment covariates based on litter size and death rate have been in use as an ad-hoc method to improve model fit, Carey establishes a theoretical basis that justifies the use of such adjustment covariates and specifies what adjustment covariates are appropriate given the assumed likelihood.

1.7 Research Plan

Analysis of developmental toxicology data presents several layers of statistical challenges, brought on by the litter effect and the correlation between multiple outcomes of interest. Early methods have focused on specific pieces of the problem, such as the litter effect, while later methods have built on these early models and added features to account for multiple outcomes and their inherent correlation. More recent methods, for example, allow an analysis which incorporates the hierarchical nature of live and non-live outcomes while not resorting to the assumption of conditional independence and still accounting for clustering.

The ultimate goal for developmental toxicity data analysis is to use the data to conduct an informed risk assessment of the toxin under study while characterizing the dose-response relationships of individual outcomes. Typically, this is done by calculating a joint BMD and associated BMDL. When multiple outcomes are concerned, a joint likelihood that fully specifies how the multiple outcomes are correlated is required to calculate a joint BMD. While GEEs may be used to ensure estimates for model parameters are robust to mis-specification of between-litter correlations, as is the case with the Regan-Catalano model, ultimately, the joint likelihood of the multiple outcomes must be specified in order to characterize total risk and calculate a joint BMD. Thus, in any method

with applications to developmental toxicology, there is a trade-off between specifying enough likelihood to allow calculation of a BMD and forgoing making such assumptions in favor of methods that calculate robust estimates that are not dependent on likelihood assumptions.

Two methods, Christensen's model and Carey's model, discussed previously incorporate death as an outcome while not resorting to assuming conditional independence. Both methods assume underlying latent variables following a normal distribution to characterize the distribution of outcomes. This general approach, which assumes malformations and deaths are observed when thresholds for latent variables are exceeded, appeals to intuition and toxicological theory. It also presents a natural way to describe inter-outcome correlation. However, these approaches have some limitations. First, the accuracy of inference and risk estimation based on these models depends on the initial likelihood assumptions. Second, the latent formulation of the likelihood makes some theoretical parameters non-estimable. In the case of Carey's model, the latter issue forces the use of approximations to the actual adjustment covariates derived from the theory for the conditional malformation model.

An alternative approach that may work to circumvent some of these issues is to explore using a Plackett-Dale-type model instead of the multivariate normal to model the three outcomes of interest. Recall that the Plackett-Dale approach for mixed outcomes discussed in this paper. For live outcomes, the association between malformation and weight is defined by the global cross-ratio:

$$\psi_k = \frac{F_{w_k,m_k}(1 - F_{w_k} - F_{m_k} + F_{w_k,m_k})}{(F_{w_k} - F_{w_k,m_k})(F_{m_k} - F_{w_k,m_k})}$$

which is used to derive the joint cumulative distribution:

$$F_{w_k,m_k} = \begin{cases} \frac{1+(F_{w_k}+F_{m_k})(\psi_k-1)-S(F_{w_k},F_{m_k},\psi_k)}{2(\psi_k-1)} & \psi_k \neq 1 \\ F_{w_k}F_{m_k} & \psi_k = 1 \end{cases} \tag{1.2}$$

where

$$S(F_{w_k}, F_{m_k}, \psi_k) = \sqrt{[1 + (\psi_k - 1)(F_{w_k} + F_{m_k})]^2 + 4\psi_k(1 - \psi_k)F_{w_k}F_{m_k}}$$

which can, in turn, be used to derive the joint density function f_{w_k,m_k}. The original approach discussed in the paper by Molenberghs *et al.* considers the association due to clustering as a nuisance and therefore the psuedo-likelhiood score function used, $pl = \sum_{k=1}^{K} \sum_{j=1}^{n_k} ln(f_{w_{jk},m_{jk}}(w,m))$, does not incorporate any parameters defining the association between littermates. Geys *et al.* do propose an extension of the log-pseudolikelihood function above that includes intra-litter association parameters. The log-pseudolikelihood has the following form:

$$pl = \sum_{k=1}^{K} \sum_{j=1}^{n_k} ln(f_1(w_{jk}, m_{jk})) + \sum_{k=1}^{K} \sum_{j \neq j'}^{n_k} ln(f_2(w_{jk}, m_{j'k})) + \sum_{k=1}^{K} \sum_{j'<j} ln(f_3(w_{jk}, w_{j'k})) + \sum_{k=1}^{K} \sum_{j'<j} ln(f_4(m_{jk}, m_{j'k}))$$

where f_1, f_2, f_3, f_4 are all bivariate Plackett densities, but characterized by different odds ratios. That is, f_1 is the joint probability of weight and malformation from the same fetus, f_2 is the joint probability of weight and malformation of two different fetuses in the same litter, f_3 is the joint probability of weights between two different animals in the same litter and f_4 is the joint probability of malformations between two different animals in the same litter. Thus, instead of using the global cross-ratio to define only the association between malformation and weight of a fetus, Geys proposes using the same Plackett framework to define all associations present within a litter. By assuming exchangeability within litters, the number of cross-ratios to be estimated is reduced to four.

Borrowing this framework of using cross-ratios to determine within-litter association may be of use in establishing a Plackett-Dale approach to developmental toxicology that includes death as an outcome. While still defining a distribution for the outcomes so that univariate and joint BMDs can be calculated, the approach allows greater flexibility in deciding marginal distributions for the outcomes. In particular, binary variables such as malformations can be modeled directly by a Bernoulli distribution rather than through a more complex latent normal distribution. This feature may allow circumventing the issue of non-estimable parameters that affects methods based on the multivariate normal distribution.

Given that malformation and death are hierarchical outcomes, it may be useful to think of death, malformation, and absence of an adverse outcome as three possible outcomes on an ordinal scale. Methods that model multiple ordinal responses based on the Plackett distribution have been developed but not adapted to clustered multiple outcome litter data, (Molenberghs and Lesaffre, 1994) and may be used as a template for a model directly applicable to developmental toxicology. In particular, certain assumptions can be exploited to reduce the number of association parameters that need to be estimated. For example, by assuming exchangeability, we can claim the association between any two littermates is identical. Thus, we should be then able to incorporate the intra-litter correlation through a bivariate density of two ordinal variables, the outcome of fetus j and the outcome of fetus j', to characterize the litter association and avoid conditional independence.

In the ordinal scale, the global cross ratios can be interpreted as cumulative odds ratios. In the case of two trinomial outcomes, this amounts to four ratios to estimate. Letting H_j, M_j, and D_j denote no adverse event, malformation, and death for fetus j, respectively, the four cross-ratios are defined to be:

$$\psi_1 = \frac{P(H_j \cup M_j | H_{j'} \cup M_{j'}) / P(D_j | H_{j'} \cup M_{j'})}{P(H_j \cup M_j | D_{j'}) / P(D_j | D_{j'})}$$

$$\psi_2 = \frac{P(D_j | H_{j'} \cup M_{j'}) / P(H_j \cup M_j | H_{j'} \cup M_{j'})}{P(D_j | D_{j'}) / P(H_j \cup M_j | D_{j'})}$$

$$\psi_3 = \frac{P(H_j \cup M_j | D_{j'}) / P(D_j | D_{j'})}{P(H_j \cup M_j | H_{j'} \cup M_{j'}) / P(D_j | H_{j'} \cup M_{j'})}$$

$$\psi_4 = \frac{P(D_j | D_{j'}) / P(H_j \cup M_j | D_{j'})}{P(D_j | H_{j'} \cup M_{j'}) / P(H_j \cup M_j | H_{j'} \cup M_{j'})}$$

Note that, because the two ordinal outcomes measure associations between are the same variable but on different fetuses, ψ_1 and ψ_4, as well as ψ_2 and ψ_3, are the same cumulative odds-ratios. Thus, number of association parameters to estimate can be reduced to two in this setting.

Given this framework, we can construct the joint distribution of death and malformation which accounts for clustering via the two cross-ratio parameters. Furthermore, we will explore how the distribution can be factorized into the resulting marginal distri-

bution for death and conditional distribution of malformation given death. It will be of particular interest to discover how the death of littermates affects the conditional distribution of malformation and, specifically, how it compares to Carey's model based on the multivariate normal distribution.

Once a method that accounts for the association between death and malformation without resorting to assuming conditional independence is developed, it will be of interest to explore ways to extend the model to include fetal weight via the Plackett-Dale approach discussed above. In particular, the conditional malformation model may be used to motivate deriving the distribution of weight and malformation conditional on death similar to (2). Again, it will be of interest to derive the conditional distribution of fetal weight and malformation in order to assess how the death of littermates may affect the joint distribution malformation and weight and whether the adjustment covariates motivated by this conditional distribution are comparable to those described by Carey and Christensen.

In both cases, it will be necessary to explore the forms of dose-response models to fit for the mean and association parameters that will depend on dose, and how to formulate the pseudolikelihood for robust correction of litter effects not captured by the Plackett-Dale formulation. Estimating equations will be derived from score functions of the log-psuedolikelihood similar to the one used in the Placket-Dale model. The use of sandwich estimators similar to (1.1) may be necessary to ensure the robustness of variance estimators. Since we are developing a likelihood model, it will be possible to use a χ^2 goodness-of-fit statistic to empirically evaluate the how well the model fits the data. Methods for assessing joint risk so that a BMD can be calculated is another issue that must be addressed. In the case where death and malformation are the only outcomes of interest, the risk of an adverse event can be characterizes by $1 - P(H)$. However, the inclusion of fetal weight as an outcome may complicate how to characterize joint risk depending on the nature of the model developed.

A comparison to already developed models that do not assume conditional inde-

pendence, namely Christensen's method and Carey's method, ideally using the same datasets, would be needed to evaluate whether the new methods give good fitting parameter estimates and BMDs. We have access to many datasets to empirically evaluate the method, including 10 EPA datasets of a variety of chemicals of standard sample size (100-150 dams per study) as well as 1 very large study of the chemical 2,4,5-T with over 10,000 dams. In addition, 150 or so other National Toxicology Program (NTP) studies in multiple agents are available that can be used to empirically evaluate "asymptotic"-like behavior. These include many positive studies as well as some negative studies, so the model can be tested on a variety of data patterns. Also, simulation studies to assess how robust the method is to deviations from the assumptions and performance under various study sample sizes will also be conducted.

A Novel Method for Modeling Hierarchical Outcomes in Developmental Toxicity Data Based on the Plackett-Dale Distribution

Frederick Prichard Cudhea

Department of Biostatistics

Harvard School of Public Health

2.1 Introduction

Controlled animal studies are used to study the effects of various potentially toxic substances such as drugs or environment contaminants. In such studies, human subjects are not appropriate and researchers must rely on animal studies to assess toxicity from experimental data. Developmental toxicology studies are designed to examine the effect of chemical substances on developing organisms. These studies involve exposing pregnant animals (usually mice, rats, or rabbits) to a test substance during pregnancy and examining the effects on the fetuses. Studies typically use three or four dose groups plus a control group, with at least 20 dams per dose group. The dams are sacrificed before delivery and the contents of the uterus examined. Outcomes of interest typically include number of resorptions (early deaths), number of fetal deaths, and out of the surviving fetuses: the number and type of malformations, fetal weights and fetal lengths. Malformations are typically categorized into three general types: Skeletal, Visceral, or External. Figure 4.1 illustrates the relationships between all the various outcomes of interest (Kimmel and Price, 1990). The outcomes given the most emphasis in determining safe doses are number of embryolethalities (resorption and deaths), number of malformations, and reductions in fetal weight.

As one can see from figure 4.1, the data involve many correlations that must modeled, making proper analysis challenging. For one, the major units of observation are clustered into litters so intra-litter correlation between outcomes from the same dam is expected. Secondly, among the live fetuses, we are interested in multiple outcomes (malformation status and fetal weight) from each fetus and an inter-outcome correlation is also expected. This correlation is usually not trivial and must be properly modeled for valid inference. Also, the fact that malformation status is a binary outcome while fetal weight is a continuous outcome adds another layer of complication. Third, the hierarchical relationship between the live outcomes and death further complicates interpretation the data. That is to say, the live outcomes (malformation status and fetal weight) may not only be correlated with other live fetuses, but also with dead fetuses within the saml litter, and this

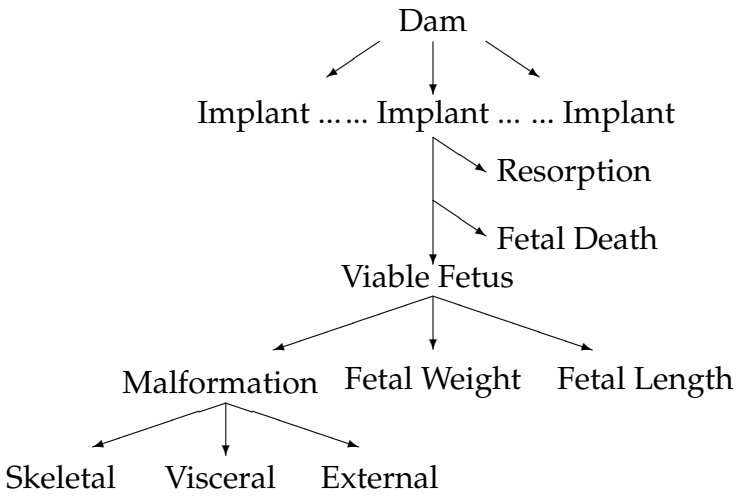

Figure 2.1: Outcomes in Developmental Toxicity

correlation should not be ignored in the data analysis.

The ultimate goal of the data analysis is to fit a dose-response model to each outcome, and to use these models to inform safe doses for regulation purposes. A key step translating the dose-response model to a 'safe' a dose is the calculation of the BMD (benchmark dose) and BMDL (benchmark dose - lower bound) (Gaylor et al., 1998), a process referred to as quantitative risk estimation, part of the larger goal of quantitative risk assessment. The BMD is defined as the dose that corresponds to a given x % increase in risk above background, where x is usually 5 or 10. The BMDL is the statistical lower-bound (usually 95%) of the BMD, and is the quantity most useful in assessing and establishing safety standards. Often, a BMDL is calculated for each outcome and the smallest is chosen, which can lead to underestimating the safe dose and ignores any correlation. A more valid approach would be to calculate a joint BMD that accounts for the combined risk of all outcomes. This approach requires that joint risk, the probability of any adverse outcome, be estimable, meaning that a joint distribution for the outcomes must be specified and that relevant inter-outcome correlations must be estimated. For methods where this is not possible (often because inter-outcome correlations are not estimated), conditional independence is assumed. That is, it is assumed that the

Table 2.1: Malformation rates by different death rates and dose for 2,4,5-T data (CD-1 strain)

	Controls and No Intervention		0.020 or 0.030 g/kg	
Death Rate	Live Fetuses Count	Malformation Rate	Live Fetuses Count	Malformation Rate
≤ 5 %	2231	0.0049	2338	0.0081
5 - 15 %	1568	0.0038	2068	0.0063
15 - 25 %	783	0.0026	1126	0.0142
25 - 35 %	85	0.0	213	0.0
35 - 45 %	41	0.0	48	0.0833
>45 %	24	0.0417	25	0.1600
	0.045 g/kg or 0.060 g/kg		0.075 or 0.090 g/kg	
Death Rate	Live Fetuses Count	Malformation Rate	Live Fetuses Count	Malformation Rate
≤ 5 %	1117	0.0858	119	0.5714
5 - 15 %	1270	0.1402	118	0.4492
15 - 25 %	1025	0.1737	107	0.7196
25 - 35 %	139	0.3237	16	1.0
35 - 45 %	70	0.2143	13	0.7692
>45 %	100	0.69	64	0.8594

live outcomes (malformation and fetal weight) are independent of the death outcomes. In other words, the death rate of a litter does not inform the malformation rate (or fetal weights) of the litter. Thus, for example, if we are only interested in death and malformation outcomes, the joint risk, $P(AdverseEvent) = P(DeadorMalformed)$, simplifies to $1-(1-P(Dead)*(1-P(Malformed|NotDead)))$. The approach, while commonly used, is not satisfying, as there is no theoretical basis for this assumption, and indeed, an examination of a large data set, from the 2,4,5-Trichlorophenoxyacetic Acid Developmental Toxicity Study (Chen and Gaylor, 1992), suggests there is a noticeable positive association between death rate and conditional malformation rate. Table 2.1 shows that for any given dose, litters with higher death rates tend to have higher conditional malformation rates as well.

2.1.1 Previous Methods

Early research focused on the problem of accounting for intra-litter correlation when only considering a single binary outcome, like embryolethality. Many important early models were developed in the late 1970's and early 1980's, including the Beta-binomial model (Williams, 1975), an extension of the binomial model, and the Ochi-Prentice model (Ochi and Prentice, 1984), which used an underlying latent multivariate normal distribution to describe the intra-litter correlation. The development of generalized estimating equations (GEE) (Liang and Zeger, 1986) allowed researchers to model this data and perform accurate inference without having to correctly specify the distributions or correlations of the outcomes, making it a popular method for analyzing not just developmental toxicity data, but a wide variety of clustered discrete data.

The research on mixed outcomes has largely focused on methods based on the latent multivariate normal distribution, which gives us an intuitive and relatively simple way to characterize the correlation between malformation and fetal weight. Catalano and Ryan (Catalano and Ryan, 1992), as well as Fitzmaurice and Laird (Fitzmaurice and Laird, 1995), take advantage of the fact that the joint likelihood can be expressed as the product of the marginal distribution for weight and conditional distribution of malformation given weight. The factorization allows weight and malformation to be modeled separately while still accounting for their correlation. Neither method is, however, conducive for formal risk estimation as joint BMDs cannot be calculated using either model. Regan and Catalano (Regan and Catalano, 1999) extend and improve on Catalano and Ryan's methodology. Their model, while still using the factorization of the latent normal distribution as a framework, allows for the estimation of the inter-outcome correlation by dose which makes the joint risk, and therefore the BMD and BMDL, possible to calculate. The correlation parameters tend to increase with dose so allowing for them to be modeled as a function of dose is an important feature of the methodology.

2.1.2 Plackett-Dale Models

Molenbergs, Geys, and Buyse (Molenberghs et al., 2001) have taken an alternative approach, using the Plackett-Dale distribution to model the two outcomes of interest. The Packett-Dale (Plackett, 1965) (Dale, 1986) approach has an advantage over more traditional probit models in that there is flexibility in choosing the marginal distributions of the outcomes. So, for example, it is possible to assume the marginal distribution for malformation is binomial rather than what is implied, for example, by the latent normal. Let $F_{w_k}(x)$ be the cumulative distribution function for w_k, the fetal weight of a fetus from litter k and let $F_{m_k}(y)$ be the cumulative distribution function for m_k, the malformation status of a fetus from litter k. Then, if (m_k, w_k) follows a Plackett-Dale distribution, their joint cumulative distribution function is

$$F_{w_k,m_k} = \begin{cases} \frac{1+(F_{w_k}+F_{m_k})(\psi_k-1)-S(F_{w_k},F_{m_k},\psi_k)}{2(\psi_k-1)} & \psi_k \neq 1 \\ F_{w_k}F_{m_k} & \psi_k = 1 \end{cases}$$

where

$$S(F_{w_k}, F_{m_k}, \psi_k) = \sqrt{[1+(\psi_k-1)(F_{w_k}+F_{m_k})]^2 + 4\psi_k(1-\psi_k)F_{w_k}F_{m_k}}$$

ψ_k, known as the global cross-ratio, defines the dependence structure of w_k and m_k,

$$\psi_k = \frac{F_{w_k,m_k}(1-F_{w_k}-F_{m_k}+F_{w_k,m_k})}{(F_{w_k}-F_{w_k,m_k})(F_{m_k}-F_{w_k,m_k})}$$

and is used to derive the above joint cumulative density function. A psuedo-likelihood based estimating equation, $pl = \sum_{k=1}^{K}\sum_{j=1}^{n_k} ln(f_{w_{jk},m_{jk}}(w,m))$, is used to estimate dose-response parameters.

Geys et al. (Geys et al., 2001) suggest, but do not implement, an extension of this method that also estimates the within-litter associations for the malformation and weight outcomes in which all associations are estimated. As we can see from figure 2.2, the within-fetus malformation-weight association is not the only source of correlation. There are also the litter-level associations, namely the association between malformation outcomes within a litter, the association between fetal weights within a litter, and the association between malformation outcomes and fetal weights between fetuses of the same

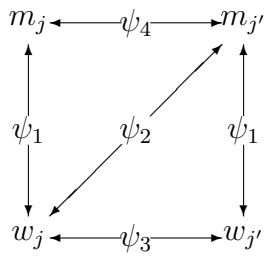

Figure 2.2: Associations present in live outcomes developmental toxicity data

litter, but these are essentially ignored in this method. One approach is to define each of these associations through a global cross-ratio and then define Plackett-Dale distributions around the cross-ratios. The estimating equation are based on a log-pseudolikelihood with the following form:

$$pl = \sum_{k=1}^{K}\sum_{j=1}^{n_k} ln(f_1(w_{jk}, m_{jk})) + \sum_{k=1}^{K}\sum_{j\neq j'}^{n_k} ln(f_2(w_{jk}, m_{j'k})) + \sum_{k=1}^{K}\sum_{j'<j} ln(f_3(w_{jk}, w_{j'k})) + \sum_{k=1}^{K}\sum_{j'<j} ln(f_4(m_{jk}, m_{j'k}))$$

where f_1, f_2, f_3, f_4 are all Plackett densities characterizing different odds ratios: f_1 is the joint probability of weight and malformation from the same fetus, f_2 is the joint probability of weight and malformation of two different fetuses in the same litter, f_3 is the joint probability of weights between two different animals in the same litter and f_4 is the joint probability of malformations between two different animals in the same litter. Note that, by assuming exchangeability within litters, the number of intra-litter association parameters to be estimated is reduced to three. Also note that, for the f_2, f_3, and f_4 parts of the log-pseudolikelihoods, we are summing over all possible pair-combinations within a litter.

2.1.3 Hierarchical Relationship Between Outcomes

Less research has been done on accounting for the correlation induced by the hierarchical relationships between death and the live outcomes (malformation and fetal weight).

For the sake of simplicity, we will ignore fetal weight and focus on the case where only death and malformation are outcomes of interest. Dose response modeling is interested in estimating the probability of death for a fetus as well as probability of malformation given the fetus survived and it is straight forward to estimate them separately within a dam. However, estimating joint risk of both outcomes is not as intuitive, unless we assume conditional independence. Because this assumption is not necessarily expected to be true in litter data, to compensate models typically include a covariate (usually litter size or proportion dead) for the malformation dose-response model to serve as an ad-hoc adjustment for the effect of death on malformation. This approach acknowledges the hierarchical nature of the data by separating the effect of dose and the effect of death-rate on malformation in the modeling. However, in joint risk assessment, this hierarchical correlation is still often ignored and conditional independence is still assumed when calculating joint risk.

Most methods proposed for this problem have been inspired from previous work relying on the latent multivariate normal distribution. Christensen (Christensen, 2004) proposes an extension to the Ochi-Prentice model, where death, malformation, and healthy outcomes are considered ordinal. Specifically, two threshold parameters, τ_m and τ_d, are used to define how the latent variable relates to the observed outcomes. Letting $\tilde{y}_{jk}$ be the latent variable for fetus j from dam k, if $\tilde{y}_{jk} < \tau_m$, then no adverse event is observed for that fetus, if $\tau_m < \tilde{y}_{jk} < \tau_d$, then a malformation is observed for the fetus and if $\tilde{y}_{jk} > \tau_d$, a fetal death is observed. $\tilde{\mathbf{y}}_k$, the vector denoting the latent variables for the fetuses from dam k, is assumed to follow a multivariate normal distribution with mean $\mu\mathbf{1}_n$ and variance $\sigma^2((1-\rho)\mathbf{I}_n + \rho\mathbf{J}_n)$. Letting H_k denote the number of healthy fetuses from litter k, M_k denote the number of malformed fetuses from litter k, and D_k denote the number of dead fetuses from litter k, the joint distribution of the three outcomes from litter k can be expressed as:

$$P(H_k, M_k, D_k) \propto \int_B \phi_n(\tilde{\mathbf{z}}_k|0, 1, \rho)d\tilde{\mathbf{z}}_k$$

where

$$\begin{aligned}
\tilde{\mathbf{z}}_k &= \tilde{\mathbf{y}}_k - \mathbf{1}_{n_k}\mu \\
B &= (\tilde{\mathbf{z}}_k | (\tilde{z}_k < -\gamma_m, k \leq H_k) \cup (\gamma_m \leq \tilde{z}_{jk} < -\gamma_d, H_k < k < H_k + M_k) \\
&\quad \cup (\tilde{z}_{jk} \geq -\gamma_d, j > H_k + M_k))) \\
\gamma_m &= \tau_m - \mu \\
\gamma_d &= \tau_d - \mu
\end{aligned}$$

γ_m and γ_d are standardized cutpoints.

Using the above likelihood, the model specification is as follows:

$$\begin{aligned}
\gamma_m &= \tau_m^*(\mathbf{X}_m\boldsymbol{\lambda}_m) + \mu(\mathbf{X}_1\boldsymbol{\beta}) \\
\gamma_D &= \tau_d^*(\mathbf{X}_d\boldsymbol{\lambda}_d) + \mu(\mathbf{X}_1\boldsymbol{\beta}) \\
\rho &= g(\mathbf{X}_2\boldsymbol{\xi})
\end{aligned}$$

where $g(\cdot)$ can either be the identity function or Fisher's Z-transformation. $\mathbf{X}_1$ is a matrix of litter-specific covariates common to both both thresholds, while $\mathbf{X}_m$ and $\mathbf{X}_d$ are litter specific litter-covariates specific to each threshhold. $\mathbf{X}_2$ is the matrix of litter-specific covariates to ρ. $\boldsymbol{\beta}$, $\boldsymbol{\lambda}_m$, $\boldsymbol{\lambda}_d$, and $\boldsymbol{\xi}$ are their respective model pararemter vectors.

Thus, since the status of a fetus is assumed to be determined by a latent normal distribution, one correlation parameter, ρ, characterizes all three correlations of interest. Estimation under certain data scenarios can be difficult. However, calculating joint risk, the risk that a fetus experiences death or a malformation, is very easy and intuitive under this model (joint risk is simply $\Phi(\gamma_m)$).

Carey (Carey, 2006) develops a simpler model that still allows for conditional dependence. Essentially, the model formalizes the ad-hoc approach of adding an adjustment covariate to the malformation dose-response model to adjust for the death-malformation correlation. The adjustment covariate is derived from a latent multivariate normal distribution. However, because the correlation parameters are reparameterized into the adjust-

ment covariate, joint risk estimation is not intuitive. Also, because the adjustment variable is based on the continuous normal distribution while the observed data are binary, the actual adjustment covariate used is an approximation of the theoretical adjustment covariate. It is unclear whether these approximations are accurate or whether it potentially introduces bias.

Because we compare our proposed method with Carey's method, we present Carey's method more formally, as applied to our situation, when only death and malformation are outcomes of interest. Unlike Christensen's model, Carey's likelihood uses two latent variables, one for death and one for malformation, denoted $\tilde{d}$ and $\tilde{m}$ respectively. The two latent variables are assumed to follow a multivariate normal distribution. More specifically, for the k-th litter:

$$\begin{pmatrix} \tilde{\mathbf{d}}_k \\ \tilde{\mathbf{m}}_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_m \end{pmatrix}, \begin{pmatrix} \Sigma_d & \Sigma_{dm} \\ \Sigma_{dm} & \Sigma_m \end{pmatrix} \right)$$

where

$$\begin{aligned}
\mu_d &= (\tilde{\alpha}_0 + \tilde{\alpha}_1 dose_k)\mathbf{1}_{n_k} \\
\mu_m &= (\tilde{\beta}_0 + \tilde{\beta}_1 dose_k)\mathbf{1}_{l_k} \\
\Sigma_d &= \sigma_d^2((1-\rho_d)\mathbf{I}_{n_k} + \rho_d \mathbf{J}_{n_k}) \\
\Sigma_m &= \sigma_m^2((1-\rho_m)\mathbf{I}_{l_k} + \rho_m \mathbf{J}_{l_k}) \\
\Sigma_{dm} &= \Sigma_{md}^T = \rho_{md}\sigma_m\sigma_d \mathbf{J}_{n_k \times l_k}
\end{aligned}$$

and l_k denotes the number of live fetuses while n_k denotes the number of implants in litter k.

Given the above likelihood, the marginal distribution of death and conditional distribution of fetal weight and malformation can be expressed as:

$$\begin{pmatrix} \tilde{\mathbf{d}}_k \\ \tilde{\mathbf{m}}_k | \tilde{\mathbf{d}}_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_{m|d} \end{pmatrix}, \begin{pmatrix} \Sigma_d & \mathbf{0}_{n_k \times l_k} & \mathbf{0}_{n_k \times l_k} \\ \mathbf{0}_{l_k \times n_k} & \mathbf{0}_{l_k} & \Sigma_{m|d} \end{pmatrix} \right)$$

where $\mu_{m|d}$ can be expressed as

$$\mu_{m|d} = (\tilde{\beta}_0 + \tilde{\beta}_1 dose) + (\rho_{md}\sigma_m)(1 + \rho_d(n_k - 1))^{-1} \left(\frac{\Sigma_{j=1}^{n_k} \tilde{d}_{ij} - n_k(\tilde{\alpha}_0 + \tilde{\alpha}_1)dose}{\sigma_d} \right).$$

Note that $\mu_{m|d}$ can be expressed as the sum of marginal model for latent malformation plus an adjustment covariate that is a function of the mean standardized residual for fetal death. While the adjustment term is a bit complicated and includes parameters from the latent theory that are not estimable, this theoretical model is used to motivate a simpler adjustment term:

$$\mu_{m|d} = (\beta_0 + \beta_1 dose) + \beta_2 \left(\frac{\bar{d}_k - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)}{\sqrt{\Phi(\hat{\alpha}_0 - \hat{\alpha}_1 dose)[1 - \Phi(\hat{\alpha_0} + \hat{\alpha}_1 dose)]/n_k}} \right)$$

Mean models are then fit using GEEs with the following dose-response framework:

$$E[d_{jk}]/\sqrt{Var(d_{jk})} = \Phi(\alpha_0 + \alpha_1 dose_k)$$
$$E[m_{jk}]/\sqrt{Var(m_{jk})} = \Phi(\beta_0 + \beta_1 dose_k)$$

To enable easy comparison between the models, we use a logit version of Carey's method rather than the proposed probit model. Given the two link functions tend to estimate similar trends in practice, we believe the derived adjustment covariate derived by Carey will still apply when the logit link function is used. Thus, in our comparisons, the Carey model was fit using the following dose-response functions:

$$\begin{aligned} logit(E[d_{jk}]) &= \alpha_0 + \alpha_1 dose_k \\ logit(E[m_{jk}|\bar{D}]) &= \beta_0 + \beta_1 dose_k + \beta_2 \left(\frac{\bar{d}_k - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)}{\sqrt{\Phi(\hat{\alpha}_0 - \hat{\alpha}_1 dose)[1 - \Phi(\hat{\alpha_0} + \hat{\alpha}_1 dose)]/n_k}} \right) \end{aligned} \quad (2.1)$$

2.2 Proposed Method

We propose a method using the Plackett-Dale framework to model dose-response for hierarchical data. It essentially takes an approach similar to Geys et al. in their proposed extension but applies it to hierarchical data. As discussed earlier, the Plackett-Dale approach has certain advantages. It is not restricted to assuming the marginal distributions

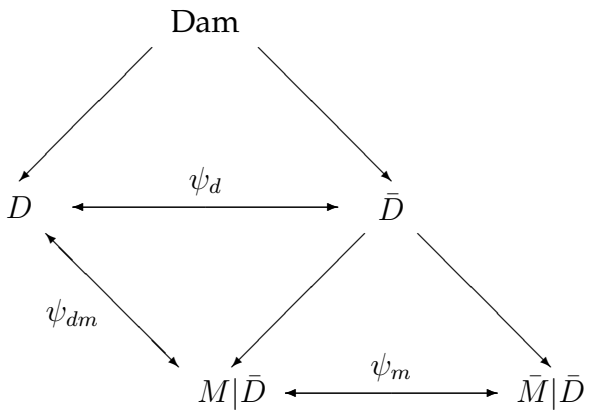

Figure 2.3: Associations present in heirarchical devlopmentoal toxicity data

are latent normal. Instead the marginal distributions are flexible. This also gives added flexibility in choosing the link function for the dose-response models and also allows separate marginal models for each outcome, rather than have to model one outcome conditional on the other. These advantages allow the resulting models to be easier to interpret. It also allows for the direct estimation and modeling of the association parameters, providing a potential path to calculating joint risk.

The various outcomes and associations of interest present within a litter can be visualized in figure 4.2. There is also the association between death outcomes within a cluster. For the fetuses that did not die, there is the association between malformation outcomes within a cluster. Finally, there is the association between death outcomes and malformation outcomes, which determines how the death rate of a particular dam will affect the corresponding conditional malformation rate for the same litter. The idea here is not to create a complicated likelihood model that will incorporate all possible outcomes and possible associations, but rather to apply the Plackett-Dale framework to each association parameter so that we can estimate all relevant association parameters as well as the death rates and conditional malformation rates.

Formalizing the notation, let d_{jk} be a binary random variable that is 1 if fetus j from dam k is dead and 0 if alive, and let ψ_d be the odds ratio of a fetus j death outcome when fetus j', a fetus in the same dam group, is also dead vs when fetus j' is not dead. Likewise,

let $m_{jk}|\bar{D}_{jk}$ be a binary random variable that is 1 if fetus j from dam k is malformed and 0 if not, given that fetus jk is known to not be dead, and let ψ_m be the odds ratio of fetus j having a malformation outcome when fetus j' is also malformed vs when fetus j' is not malformed, assuming both fetus j and j' are not dead. Finally, let ψ_3 be the odds ratio of fetus j (which is known to be alive) having a malformation outcome when fetus j' is dead vs when fetus j' is not dead. Mathematically, their expressions are as follows:

$$\psi_d = \frac{P(D_j|D_{j'})/P(\bar{D}_j|D_{j'})}{P(D_j|\bar{D}_{j'})/P(\bar{D}_j|\bar{D}_{j'})}$$

$$\psi_m = \frac{P\left(M_j|\bar{D}_j\middle|M_{j'}|\bar{D}_{j'}\right)/P\left(\bar{M}_j|\bar{D}_j\middle|M_{j'}|D_{j'}\right)}{P\left(M_j|\bar{D}_j\middle|\bar{M}_{j'}|\bar{D}_{j'}\right)/P\left(\bar{M}_j|\bar{D}_j\middle|\bar{M}_{j'}|\bar{D}_{j'}\right)}$$

$$\psi_{dm} = \frac{P\left(M_j|\bar{D}_j\middle|D_{j'}\right)/P\left(\bar{M}_j|\bar{D}_j\middle|D_{j'}\right)}{P\left(M_j|\bar{D}_j\middle|\bar{D}_{j'}\right)/P\left(\bar{M}_j|\bar{D}_j\middle|\bar{D}_{j'}\right)}$$

where D_j is a death outcome for fetus j and M_j is a malformation outcome for fetus j. As with Regan-Catalano's method, a dose-response model can be estimated for the association parameter, ψ, and thus BMDs can be calculated.

Parameters ψ_d, ψ_m, ψ_{dm} can be thought of as global cross-ratios that define the various associations present in the data: ψ_d is the within-cluster association between death outcomes, ψ_m is the within-cluster association between malformation outcomes, and ψ_3 is the association between death outcome and malformation outcome that is induced by conditional dependence. From these cross-ratios, the joint probabilities for two deaths, two malformations (given they are not dead), and one death and one malformation (given the malformed fetus was known not to be dead), can be derived as:

$$F_1 = P(D_j, D_{j'}) = \begin{cases} \frac{1+(2p_d)(\psi_1-1)-S(p_d,p_d,\psi_d)}{2(\psi_d-1)} & \psi_d \neq 1 \\ p_d^2 & \psi_d = 1 \end{cases}$$

$$F_2 = P(M_j|\bar{D}_j, M_{j'}|\bar{D}_{j'}) = \begin{cases} \frac{1+(2p_{m|\bar{D}})(\psi_m-1)-S(p_{m|\bar{D}},p_{m|\bar{D}},\psi_m)}{2(\psi_m-1)} & \psi_m \neq 1 \\ p_{m|\bar{D}}^2 & \psi_m = 1 \end{cases}$$

$$F_3 = P(M_j|\bar{D}_{j'}, D_{j'}) = \begin{cases} \frac{1+(p_{m|\bar{D}}+p_d)(\psi_{dm}-1)-S(p_m|\bar{D},p_d,\psi_{dm})}{2(\psi_{dm}-1)} & \psi_{dm} \neq 1 \\ p_{m|\bar{D}}p_d & \psi_{dm} = 1 \end{cases}$$

where $S(p_1, p_2, \psi) = \sqrt{[1+(\psi-1)(p_1+p_2)]^2 + 4\psi(1-\psi)p_1p_2}$.

From these joint probabilities we can derive the the probability mass functions for the paired outcomes to be:

$$G_1(d_j, d_{j'}) = \begin{cases} F_1(p_d, \psi_d) & d_j = 1, d_{j'} = 1 \\ 2(p_d - F_1(p_d, \psi_d)) & d_j \neq d_{j'} \\ 1 - 2p_d + F_1(p_d, \psi_d) & d_j = 0, d_{j'} = 0 \end{cases}$$

$$G_2(m|\bar{D}_j, m|\bar{D}_{j'}) = \begin{cases} F_2(p_{m|\bar{D}}, \psi_m) & m|\bar{D}_j = 1, m|\bar{D}_{j'} = 1 \\ 2(p_{m|\bar{D}} - F_2(p_{m|\bar{D}}, \psi)) & m|\bar{D}_j \neq m|\bar{D}_{j'} \\ 1 - 2p_{m|\bar{D}} + F_2(p_{m|\bar{D}}, \psi_m) & m|\bar{D}_j = 0, m|\bar{D}_{j'} = 0 \end{cases}$$

$$G_3(m|\bar{D}_j, d_{j'}) = \begin{cases} F_3(p_{m|\bar{D}}, p_d, \psi_{dm}) & m|\bar{D}_j = 1, d_{j'} = 1 \\ p_{m|\bar{D}} - F_3(p_{m|\bar{D}}, \psi_{dm}) & m|D_j = 1, d_{j'} = 0 \\ p_D - F_3(p_D, \psi_{dm}) & m|D_j = 0, d_{j'} = 1 \\ 1 - p_{m|\bar{D}} - p_d + F_3(p_{m|\bar{D}}, \psi_{dm}) & m|\bar{D}_j = 0, d_{j'} = 0 \end{cases}$$

We take a similar approach to the one proposed in Geys et al. to derive the estimating equations. Geys et al. propose summing all likelihoods to create a psuedolikelihood to form the estimating equations. However, they were only considering live outcomes (malformation and fetal weight). If we take the same approach and use the following pseudolikelihood:

$$\begin{aligned} pl = &\sum_{k=1}^{K} \sum_{j'<j} ln(G_1(d_{jk}, d_{j'k})) \\ &+ \sum_{k=1}^{K} \sum_{j'<j} ln(G_2(m_{jk}|\bar{D}_{jk}, m_{j'k}|\bar{D}_{j'k})) \\ &+ \sum_{k=1}^{K} \sum_{j \neq j'} ln(G_3(d_{jk}, m_{j'k}|\bar{D}_{j'k})) \end{aligned}$$

then we ignore the hierarchical relationship inherent in the data and we get biased estimates. Simulation studies show that p_d was consistently underestimated and $p_{m|\bar{D}}$ was consistently overestimated. We believe that simultaneously estimating the parameters for the death and malformation models in this way possibly leads to a positive feedback loop due to the presence of the G_3 portion of the likelihood. The presence of G_3 portion in the pseudolikelihood means that in the estimation procedure, the estimate for $p_{m|\bar{D}}$ informs the estimate of p_d and vice versa in a very direct way and which may potentially lead to extreme bias.

Thus, we propose a 2-step procedure for the estimation. First, estimate dose response

parameters for p_d and ψ_d from estimating equations based on

$$pl_1 = \sum_{k=1}^{K} \sum_{j'<j} ln(G_1(d_{jk}, d_{j'k}))$$

Then, estimates for p_m, ψ_m, and ψ_3 can then be estimated from estimating equations based on

$$pl_2 = \sum_{k=1}^{K} \sum_{j'<j} ln(G_2(m_{jk}|\bar{D}_{jk}, m_{j'k}|\bar{D}_{j'k})) + \sum_{k=1}^{K} \sum_{j \neq j'} ln(G_3(d_{jk}, m_{j'k}|\bar{D}_{j'k}))$$

by substituting parameters for p_d and ψ_d with their estimates obtained from step one.

Thus, we estimate two dose-response models:

$$\eta_{k1} = \begin{pmatrix} logit(p_{d_k}) \\ log(\psi_{d_k}) \end{pmatrix} = \mathbf{X}_{k1}\boldsymbol{\beta}_1$$

$$\eta_{k2} = \begin{pmatrix} logit(p_{m|\bar{d}_k}) \\ log(\psi_{m_k}) \\ log(\psi_{dm_k}) \end{pmatrix} = \mathbf{X}_{k2}\boldsymbol{\beta}_2$$

We use the logit-link for the probability models and the log-link for the ψ models, but other options, such as the probit-link for the probability models are also possible.

The estimating equations used to estimate β_1 and β_2 are

$$U(\beta_1) = \sum_{k=1}^{N} \left(\frac{\partial \eta_{k1}}{\partial \beta_1}\right)^T \left(\frac{\partial \eta_{k1}}{\partial \beta_1}\right)^{-T} \left(\frac{\partial pl_1}{\partial \theta_{k1}}\right)$$

and

$$U(\beta_2) = \sum_{k=1}^{N} \left(\frac{\partial \eta_{k2}}{\partial \beta_2}\right)^T \left(\frac{\partial \eta_{k2}}{\partial \beta_2}\right)^{-T} \left(\frac{\partial pl_2}{\partial \theta_{k2}}\right)$$

respectively, where $\theta_1 = (p_d, \psi_d)$ and $\theta_2 = (p_{m|\bar{D}}, \psi_m, \psi_{dm})$.

The covariance estimates for β_1 and β_2 are

$$\hat{cov}(\hat{\beta}_1) = \left(\sum_{k=1}^{N} \frac{\partial U_k(\beta_1)}{\partial \beta_1}\right)^{-1} \left(\sum_{k=1}^{N} U_k(\beta_1) U_k(\beta_1)^T\right) \left(\sum_{k=1}^{N} \frac{\partial U_k(\beta_1)}{\partial \beta_1}\right)^{-T} \Bigg|_{\beta_1 = \hat{\beta}_1}$$

and

$$\hat{cov}(\hat{\beta_2}) = \left(\sum_{k=1}^{N} \frac{\partial U_k(\beta_2)}{\partial \beta_2}\right)^{-1} \left(\sum_{k=1}^{N} U_k(\beta_2) U_k(\beta_2)^T\right) \left(\sum_{k=1}^{N} \frac{\partial U_k(\beta_2)}{\partial \beta_2}\right)^{-T} \Bigg|_{\beta_2=\hat{\beta_2}}$$

respectively.

Because this method treats each possible pairing within a dam as the outcome, it weighs each dam differently than other methods that use the implant or fetus as an outcome. For example, when modeling p_d, a logistic regression would weight each dam by the number of implants for that dam, n_k. However, the Plackett-Dale model, because it treats each possible pairing from a dam as a data point, weights each dam by something closer to n_k^2. This can lead to potentially biased estimates and inference inconsistent with established methods. Indeed, examining the three log-likelihoods shows us that they do not simplify to what one would expect under independence. The three log-likelihoods of dam k under independence are as follows:

$$\sum_{j'<j} ln(G_1(d_{jk}, d_{j'k})) = (n_k - 1)[n_{k,d=1} ln(p_{dk}) + n_{k,d=0} ln(1 - p_{dk})]$$

$$\sum_{j'<j} ln(G_2(m_{jk}|\bar{D}_{jk}, m_{j'k}|\bar{D}_{j'k})) = (l_k - 1)[l_{k,m=1} ln(p_{m|\bar{d}}) + l_{k,m=0} ln(1 - p_{m|\bar{d}})]$$

$$\sum_{j \neq j'} ln(G_3(d_{jk}, m_{j'k}|\bar{D}_{j'k})) = (l_k) n_{k,d=1} ln(p_{dk}) + (l_k - 1) n_{k,d=0} ln(1 - p_{dk})$$
$$+ (n_k - 1)[l_{k,d=1} ln(p_{m|\bar{d}}) + l_{k,d=0} ln(1 - p_{m|\bar{d}})]$$

The likelihoods for each dam are weighted by a function of number of implants or litter size (or both) when no litter-level associations exist ($\psi_d = 1, \psi_m = 1, \psi_{dm} = 1$). Ideally, in the case that no litter-level associations exist, we expect the log-likelihood to be similar to the log-likelihoods used in a ordinary logistic regression, where each fetus is weighted equally. However, dams with larger implant sizes and litter sizes are weighted more heavily in our method, giving weight to fetuses of lower doses, which could potentially influence parameter estimates and inference. In order to prevent this extraneous weighting where dams with larger implant counts or litter sizes are weighted disproportionately higher, we divide the derivative of the log-likelihood by the weighting factor so that our

method conforms to the same weighting as a standard logistic regression under complete independence. Thus, instead of using

$$\left(\frac{\partial pl_1}{\partial \theta_{k1}}\right) = \begin{pmatrix} \sum_{j'<j} \frac{\partial}{\partial p_{d_k}} ln(G_1(d_{jk}, d_{j'k})) \\ \sum_{j'<j} \frac{\partial}{\partial \psi_{d_k}} ln(G_1(d_{jk}, d_{j'k})) \end{pmatrix}$$

and

$$\left(\frac{\partial pl_2}{\partial \theta_{k2}}\right) = \begin{pmatrix} \sum_{j'<j} \frac{\partial}{\partial p_{m|\bar{D}_k}} ln(G_2(m|\bar{D}_{jk}, m|\bar{D}_{j'k})) + \sum_{j\neq j'} \frac{\partial}{\partial p_{m|\bar{D}_k}} ln(G_3(d_{jk}, m|\bar{D}_{j'k})) \\ \sum_{j'<j} \frac{\partial}{\partial \psi_{m|\bar{D}_k}} ln(G_2(m|\bar{D}_{jk}, m|\bar{D}_{j'k})) \\ \sum_{j\neq j'} \frac{\partial}{\partial \psi_k} ln(G_3(d_{jk}, m|\bar{D}_{j'k})) \end{pmatrix}$$

we use

$$\left(\frac{\partial pl_1}{\partial \theta_{k1}}\right) = \begin{pmatrix} \frac{1}{n_k - 1} \sum_{j'<j} \frac{\partial}{\partial p_{d_k}} ln(G_1(d_{jk}, d_{j'k})) \\ \\ \sum_{j'<j} \frac{\partial}{\partial \psi_{d_k}} ln(G_1(d_{jk}, d_{j'k})) \end{pmatrix}$$

and

$$\left(\frac{\partial pl_2}{\partial \theta_{k2}}\right) = \begin{pmatrix} \frac{1}{l_k - 1} \sum_{j'<j} \frac{\partial}{\partial p_{m|\bar{D}_k}} ln(G_2(m|\bar{D}_{jk}, m|\bar{D}_{j'k})) + \frac{1}{n_k - 1} \sum_{j\neq j'} \frac{\partial}{\partial p_{m|\bar{d}_k}} ln(G_3(d_{jk}, m|\bar{d}_{j'k})) \\ \sum_{j'<j} \frac{\partial}{\partial \psi_{m|\bar{d}_k}} ln(G_2(m|\bar{D}_{jk}, m|\bar{D}_{j'k})) \\ \sum_{j\neq j'} \frac{\partial}{\partial \psi_k} ln(G_3(d_{jk}, m|\bar{D}_{j'k})) \end{pmatrix}$$

in our estimating equations.

Because the units are the paired outcomes, if no pairs exist for a particular dam, that data can't be included in the analysis. This includes litters with only one implant and with only one surviving fetus. The former case is so rare that it doesn't merit any consideration. The second case though is not impossible if the data set considered is large enough and the number of implants are relatively small. In such cases, the death outcome pairs and the malformation-death pairs can contribute to the estimation for the models for p_d, ψ_d, ψ_{dm}, but we cannot use the outcome of the surviving fetus to inform the models for $p_{m|\bar{D}}$ and ψ_m. Typically, however, this phenomenon is rarely observed in practice. The only realistic scenario in which we would observe a significant number of litters with only one surviving fetus is the case where the highest dose group has an extremely high death rate such that many of the dams in that dose group have no or just one surviving fetuses. In such situations, it is common to drop that dose-group entirely from the analysis because it will potentially affect the dose-response model greatly even though it does not really

inform the low-dose effect of the study agent due to a differing mechanism of action toxicologically at the highest dose. So for the purposes of developmental toxicity studies, this limitation is not a significant problem.

The above estimation procedure was programmed in R (version 2.15). For both steps of the estimation procedure, the Newton-Raphson algorithm for non-linear sets of equations is used to estimate the model parameters. Both backtracking and the perturbation of the jacobian when not positive definite are implemented in the algorithm (Press et al., 2007). Functions to calculate first and second derivatives for G_1, G_2, and G_3 with respect to each parameter were created. Starting values for the p_d and $p_{m\bar{d}}$ models are calculated by running the equivalent logistic regression models with GEEs. For the starting values for the ψ_d, ψ_m, and ψ_{dm} models, a starting value of 0.0001 is used. For the EG mice dataset (presented below), a study with a typical sample size and data-pattern, the first estimation procedure took 4 iterations and the second estimation procedure took 3 iterations for convergence. The first procedure took 13.36 seconds while the second procedure took 28.48 seconds on a Dell Precision 390 desktop computer with an Intel(R) Pentium(R) 4 CPU 3.00 GHz 2.99 GHz processer running Windows 7 Professional.

2.3 Example

2.3.1 NTP Study of EG in Mice

To illustrate our method, we fit the model to a developmental toxicity study conducted by the National Toxicology Program to study the effect of Ethylene Glycol (EG) in mice. Relevant summary statistics for the data set are presented in table 2.3.1. While developmental toxicity studies typically examine multiple types of malformations, for our purposes we will define a malformation outcome as any type of malformation observed. The data set exemplifies typical traits of such investigations. Both the death rate and the conditional malformation rate increase with dose, but the conditional malformation rate appears to be more sensitive to dose. In the addition, the outcomes have very different background

Table 2.2: Summary Statistics for EG Mice Data

dose (g/g/day)	Dams	Implants	Deaths	%	Malformations	%
0	25	384	37	11.1	1	0.337
0.75	24	310	34	11.0	26	9.42
1.5	23	266	37	13.9	89	38.9
3.0	23	283	57	20.4	129	57.1

Table 2.3: Model Selection for EG data for p_d and ψ_d models. An asterisk under 1, d or d^2 indicate a constant, linear, or quadratic dose trend respectively

	$logit(p_d)$			$log(\psi_d)$	
Model	1	d	d^2	1	d
1	*	*	*	*	*
2	*	*		*	*
3	*	*		*	
4	*			*	

Comparison	Wald-test statistic	p-value
1-2	0.213	0.644
2-3	0.0386	0.844
3-4	6.80	0.009

rates of response; the baseline malformation rate is near 0 but the control death rate is non-trivial at greater than 10%.

We fit the following models:

$$\begin{aligned} logit(p_d) &= \beta_{d_0} + \beta_{d_1} dose \\ ln(\psi_d) &= \alpha_{d_0} \\ logit(p_{m|\bar{d}}) &= \beta_{m_0} + \beta_{m_1} dose + \beta_{m_2} dose^2 \\ ln(\psi_m) &= \alpha_{m_0} \\ ln(\psi_{dm}) &= \alpha_{dm_0} \end{aligned} \tag{2.2}$$

We decided this model fit the data best after model selection was performed. There was not strong evidence of the association parameters changing with dose. Hence, for the sake of parsimony, they are assumed to be constant across dose groups. The quadratic model for the conditional malformation outcomes fit the data better than the linear model and thus was chosen for this analysis. Table 2.3.1 shows Wald test comparisons for the p_d and ψ_d models and Table 2.3.1 show Wald test comparisons for the p_m, ψ_m, and ψ_{dm} models.

The resulting parameter estimates and associated standard errors are shown in Table

Table 2.4: Model Selection for EG data for p_m, ψ_m and psi_{dm} models. An asterisk under 1, d or d^2 indicate a constant, linear, or quadratic dose trend respectively. All models are fit with the same death model specifications (linear in $logit(p_d)$ and constant in $ln(\psi_d)$)

	$logit(p_m)$			$log(\psi_m)$		$log(\psi_{dm})$	
Model	1	d	d^2	1	d	1	d
1	*	*	*	*	*	*	*
2	*	*		*	*	*	*
3	*	*	*	*		*	

Comparison	Wald-test statistic	p-value
1-2	17.4	0.00003
1-3	3.55	0.169

Table 2.5: Parameter Estimates, Standard Errors, and 95% Confidence Intervals for EG mice data (model 4.4)

param	estimate	standard error	95% confidence interval
β_{d_0}	-2.20	0.180	(-2.55, -1.85)
β_{d_1}	0.264	0.101	(0.07, 0.46)
α_{d_0}	0.521	0.139	(0.25, 0.79)
β_{m_0}	-5.26	0.563	(-6.36, -4.16)
β_{m_1}	4.60	0.804	(3.02, 6.18)
β_{m_2}	-0.917	0.219	(-1.35, -0.49)
α_{m_0}	1.23	0.219	(0.80, 1.66)
α_{dm_0}	0.218	0.158	(-0.09, 0.528)

A.1. Typically, the second-order parameters (here, the αs) tend to have larger standard errors and much wider confidence intervals and it is harder to achieve statistical significance for those parameters. For this data set, we see that none of the dose-response parameters for the ψ models were statistically significant at the .05 level, and in fact, the death-malformation association is not detectable at the 0.05 level for this data set (p-value = 0.17). However, the positive estimate for the parameter hints that the model is capturing the expected positive correlation between death and malformation.

The estimated probabilities for each dose group are shown in Table 2.3.1 with the estimated ψs shown in table 2.3.1.

For the death dose-reponse model, we calculated a $BMD_{0.05}$ of 1.60 g/kg/day and $BMDL_{0.05}$ of 1.10 g/kg/day. For the malformation dose-response model, we calculated a $BMD_{0.05}$ of 0.596 g/kg/day and $BMDL_{0.05}$ of 0.522 g/kg/day. As expected, the $BMD_{0.05}$

Table 2.6: Estimates for Probabilities by Dose

dose (g/kg)	LCL	$\hat{p}_d$	UCL		LCL	$log(\hat{p}_m)$	UCL
0.0	0.0726	0.100	0.136		0.00172	0.00516	0.0154
0.75	0.0951	0.120	0.149		0.0631	0.0889	0.124
1.5	0.117	0.142	0.171		0.286	0.395	0.515
3.0	0.143	0.197	0.267		0.432	0.570	0.698

Table 2.7: Estimates for ψs for EG data

ψ	estimate	95% confidence interval
ψ_d	1.68	(1.28, 2.21)
ψ_m	3.43	(2.23, 5.25)
ψ_{dm}	1.24	(0.911, 1.70)

and $BMDL_{0.05}$ estimates are much lower for malformation outcomes than death outcomes, since malformation rate was shown to be more sensitive to dose in this study.

We do not present the results of the models fitted, but we note that the probability models were very robust to how the ψ parameters were estimated. The ψ model estimates, on the other hand, were more sensitive to how the probability models were specified.

2.3.2 2,4,5-T Study in Mice

We also evaluated our method on a larger data set from a study examining the effects of 2,4,5-Trichlorophenoxyacetic Acid (2,4,5-T) (Chen and Gaylor, 1992). Several strains of mice were used in this large experiment and we examine here a subset of the data in the CD-1 strain. Relevant summary statistics for the data set are presented in table 2.3.2. As we can see by the number of dams per dose group, the study did employ a balanced study design because doses were added dynamically as the experiment was conducted over time.

In this large data set, we would expect the phenomenon of conditional dependence to be more easily observed. Christiansen made a convincing case that a correlation between conditional malformation rate and death rate for this data by looking at the conditional

Table 2.8: Malformation rates by different death rates and dose for 2,4,5-T data (CD-1 strain)

dose (g/kg/day)	Dams	Implants	Deaths	%	Malf	%
0.000	698	8061	820	10.2	33	0.456
0.020	307	3637	410	11.3	24	0.744
0.030	722	8300	1079	13.0	79	1.09
0.045	98	1120	229	20.4	110	12.3
0.060	592	6865	1408	20.5	858	15.7
0.075	44	482	214	44.4	159	59.3
0.090	83	917	494	53.9	268	63.4

malformation rates by dose and by death rate. He observes that even for dams at the same dose level, dams with higher death rates tend to also have higher malformation rates. Table 2.1 shows the malformation rates by death rates for dose group. The data from this table only includes litters with 10 to 13 implants to control for the possible effect of litter size. It also only includes litters with at least one live outcome.

Thus, for these data, we would expect to be able to detect statistically significance for the ψ_{dm} parameters. Using the same model-fitting strategy we employed for the EG data, we found the following best fitting model:

$$\begin{aligned} logit(p_d) &= \beta_{d_0} + \beta_{d_1} dose + \beta_{d_2} dose^2 \\ ln(\psi_d) &= \alpha_{d_0} + \alpha_{d_1} dose \\ logit(p_{m|\bar{d}}) &= \beta_{m_0} + \beta_{m_1} dose \\ ln(\psi_m) &= \alpha_{m_0} + \alpha_{m_1} dose \\ ln(\psi_{dm}) &= \alpha_{dm_0} \end{aligned} \tag{2.3}$$

Tables 2.3.2 and 2.3.2 show the relevant Wald-test comparisons.

We note that the baseline ψ_{dm} is indeed statistically significant. The model parameter estimates, standard errors, and 95% confidence intervals are shown in Table A.1. For this data set, we observe that α_{dm_0} is statistically significant at the two-sided 0.05 level, confirming that the death-malformation association is indeed detectable. There is no ev-

Table 2.9: Model Selection for 2,4,5-T data for p_d and ψ_d models. An asterisk under 1, d or d^2 indicate a constant, linear, or quadratic dose trend respectively

	$logit(p_d)$			$log(\psi_d)$		
Model	1	d	d^2	1	d	d^2
1	*	*	*	*	*	*
2	*	*	*	*	*	
3	*	*		*	*	
4	*	*	*	*		

Comparison	Wald-test statistic	p-value
1-2	2.10	0.147
2-3	36.01	< 0.0001
2-4	30.71	< 0.0001

Table 2.10: Model Selection for 2,4,5-T data for p_m, ψ_m and ψ_{dm} models. An asterisk under 1, d or d^2 indicate a constant, linear, or quadratic dose trend respectively. All models are fit with the same death model specifications (quadratic in $logit(p_d)$ and linear in $ln(\psi_d)$)

	$logit(p_m)$			$log(\psi_m)$			$log(\psi_{dm})$		
Model	1	d	d^2	1	d	d^2	1	d	d^2
0	*	*	*	*	*		*	*	*
1	*	*	*	*	*	*	*	*	
2	*	*	*	*	*		*	*	
3	*	*		*	*		*	*	
4	*	*		*	*		*		
5	*			*	*		*		
6	*	*		*			*		

Comparison	Wald-test statistic	p-value
1-2	2.72	0.100
0-2	1.73	0.188
2-3	1.12	0.290
3-4	0.087	0.768
4-5	671.7	< 0.0001
4-6	9.74	0.00180

Table 2.11: Parameter Estimates, Standard Errors, and 95% Confidence Intervals for 245T data (CD-1 strain)

param	estimate	standard error	95% confidence interval
β_{d_0}	-2.15	0.0551	(-2.26, -2.04)
β_{d_1}	-3.58	3.75	(-10.9, 3.77)
β_{d_2}	304.23	50.69	(204.9, 403.6)
α_{d_0}	0.887	0.149	(0.596, 1.19)
α_{d_1}	16.7	3.01	(10.8, 22.6)
β_{m_0}	-6.33	0.174	(-6.68,-5.99)
β_{m_1}	79.3	3.06	(73.3, 85.3)
α_{m_0}	3.51	0.367	(2.79, 4.23)
α_{m_1}	-18.6	5.94	(-30.2, -6.90)
α_{dm_0}	0.613	0.0739	(0.468, 0.758)

idence that this association increases with dose. The significantly positive α_{dm_0} confirms that conditional dependence is detectable with a sufficiently large data set. We also note that the method estimates α_{m_1} to be negative and the baseline association to be extremely high. It is possible this is an artifact of the low malformation rates in the control group. When an extreme number of the same outcome (in this case, non-malformed fetuses) are observed, the level of association may be inflated, resulting in a high baseline estimate for ψ_m and an estimated negative trend.

The estimated probabilities and ψ values for each dose group are shown in Table 2.3.2. The estimate for $log(\psi_{dm})$ is 0.613 with a confidence interval of (0.468, 0.758).

For the death dose-repose model, we calculated a BMD of 0.0430 g/kg/day and $BMDL_{0.05}$ of 0.0379 g/kg/day. For the malformation dose-response model, we calculated a BMD of 0.0432 g/kg/day and $BMDL_{0.05}$ of 0.0419 g/kg/day. We note that, unlike the EG data where the malformation BMD is much lower than the death BMD, the two BMDs estimates for the 2,4,5-T data are very similar. Here, the common practice of choosing the smaller of the two BMDLs as a guide to a safe dose may actually significantly underestimate the overall risk of a negative outcome. In this case, that approach would completely ignore the risk of malformation even though the malformation risk is similar to the death risk at low dose levels. This data set underlies the importance of

Table 2.12: Estimates for Probabilities and ψs by Dose

dose (g/kg)	LCL	$\hat{p}_d$	UCL		LCL	$log(\hat{\psi}_d)$	UCL
0.0	0.0946	0.104	0.115		0.596	0.887	1.18
0.020	0.102	0.109	0.117		1.02	1.22	1.42
0.030	0.112	0.121	0.131		1.21	1.39	1.56
0.045	0.144	0.155	0.167		1.48	1.64	1.80
0.060	0.203	0.219	0.236		1.69	1.89	2.08
0.075	0.297	0.330	0.366		1.88	2.14	2.39
0.090	0.428	0.498	0.569		2.06	2.39	2.72
dose (g/kg)	LCL	$\hat{p}_{m\|\bar{d}}$	UCL		LCL	$log(\hat{\psi}_{m\|\bar{d}})$	UCL
0.0	0.00126	0.00177	0.00249		2.79	3.51	4.23
0.020	0.00683	0.00860	0.0108		2.64	3.14	3.65
0.030	0.0157	0.0188	0.0225		2.56	2.96	3.36
0.045	0.0525	0.0593	0.0669		2.41	2.68	2.95
0.060	0.155	0.171	0.190		2.18	2.40	2.62
0.075	0.364	0.405	0.448		1.84	2.12	2.40
0.090	0.636	0.691	0.742		1.42	1.84	2.27

exploring satisfactory ways of calculating joint risk so that a joint BMD can be calculated for risk assessment. The proposed method does not have a straight forward formula to calculate joint risk because it is not based on a full likelihood model; a possible avenue of research is to construct a method for joint BMD calculation that takes advantage of direct calculations of the association parameter between death and malformation outcomes.

2.4 Comparison to other models

In this section, we compare the P-D model proposed here to previously proposed methods that account for correlation between death and malformation. We examine estimates for p_d and $p_{m|d}$ by dose group for four methods: the P-D method, Carey's method, a "naive" method, and an "empirical" method. The "naive" method is merely fitting a logistic regression for the death data and the malformation data with GEEs and assuming conditional independence. For Carey's model, we fit the same dose response models but add the adjustment covariate described in (2.1) for the malformation model. In both cases, we assume exchangeability for the working correlation matrix. The "empirical"

Table 2.13: Estimated Probabilities by Dose and Estimation Method for EG Mice Data

dose (g/kg)	$\hat{p}_d$-Emp	$\hat{p}_d$-Naive	$\hat{p}_d$-P-D	$\hat{p}_d$-Carey
0.0	0.108	0.102	0.100	0.102
0.75	0.110	0.122	0.120	0.122
1.5	0.139	0.145	0.142	0.145
2.0	0.201	0.202	0.197	0.202
dose (g/kg)	$\hat{p}_{m\|\bar{d}}$-Emp	$\hat{p}_{m\|\bar{d}}$-Naive	$\hat{p}_{m\|\bar{d}}$-P-D	$\hat{p}_{m\|\bar{d}}$-Carey
0.0	0.00337	0.00537	0.00537	0.00527
0.75	0.0942	0.0906	0.0889	0.0918
1.5	0.389	0.396	0.395	0.404
2.0	0.571	0.567	0.570	0.572

method simply calculates the empirical probabilities without any modeling. The empirical method gives estimates closest to the true probabilities and can be used to gauge bias that might be introduced in model-fitting for each of the methods presented.

We also compare BMD and BMDLs calculated using our method, Carey's method and the "naive" method. There are several choices for how to calculate a malformation BMD from Carey's model. We take the simplest approach by calculating risk at the mean of the residual terms. Theoretically, the mean adjustment covariates, which are based on residuals of the death model, should have mean zero. Thus, it is reasonable to interpret the dose-response parameter as the marginal dose effect on malformation (conditional on the fetus being alive). The BMDs and corresponding BMDLs are for a 5% increase in risk. Note that for the death model, Carey's method and the "naive" method are identical so a separate BMD for the naive method is not included. BMDLs are calculating using the method proposed by Kimmel and Gaylor (Kimmel and Gaylor, 1988).

2.4.1 Comparisons using EG Study Data

Table 2.4.1 shows the probability estimates by dose and estimation method. We note that both p_d and $p_{m|\bar{D}}$ estimates are fairly similar by dose across estimation methods, and are consistent with the empirical estimates.

Table 2.14: Comparison of BMD and BMDL from various models

	Death model		Malformation model		
	P-D	Carey	P-D	Carey	Naive
BMD	1.60	1.55	0.596	0.587	0.590
BMDL	1.10	1.08	0.522	0.509	0.513
Relative Difference	0.312	0.303	0.124	0.132	0.131

Table 2.14 show the BMD and BMDL estimates for each method. We see that for both the death and malformation models, our model predicts slightly higher BMD and BMDL. The differences for the death model are, however, very small. Figure 2.4 show the dose response models for death and malformation, as well as the empirical death and malformation rates by dose (shown in green). We see that both methods provide very similar fits. We also note that, while the quadratic malformation model gives us a better fit than the linear model, with probability estimates more consistent with the empirical estimates, it also assumes the malformation rate peaks at around 2.5 g/kg and then starts to decline. In other words, it is a non-monotonic trend. This may be controversial but we believe that for the purposes of benchmark dose analysis, accuracy in low dose estimates, which the quadratic model provides, is more important.

Table 2.14 also gives the relative difference between BMD and BMDLs ($\frac{BMD-BMDL}{BMD}$). We see no large difference between methods, suggesting that the BMD variability is essentialy the same for all methods.

2.4.2 Comparisons using 2,4,5-T Study Data

Table 2.4.2 shows the probability estimates by dose and estimation method. We note that the p_d estimates are fairly similar by dose across estimation methods, and appear consistent with the empirical estimates. However, we note that the $p_{m|\bar{D}}$ estimates for Carey's method and our method diverge in the higher dose groups. The $p_{m|\bar{D}}$ estimates from Carey's method is much higher than the $p_{m|\bar{D}}$ estimate from our method or the naive method for dose group (0.60 g/kg) and higher. Figure 2.5 shows the dose re-

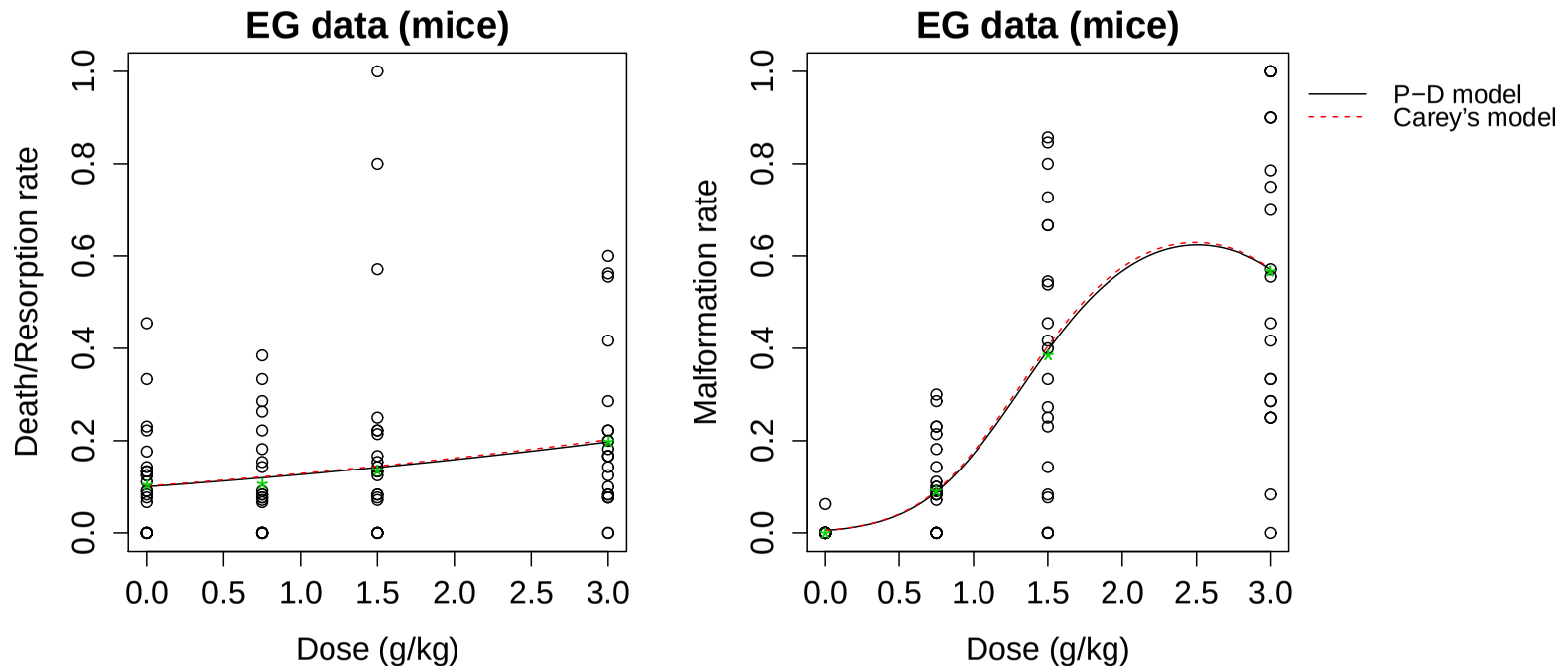

Figure 2.4: Dose Response model for death and malformation for EG data (mice). Each dot represents the observed death/resorption rate or malformation rate of a litter in the data set.

Table 2.15: Estimated Probabilities by Dose and Estimation Method

dose (g/kg)	$\hat{p}_d$-Emp	$\hat{p}_d$-Naive	$\hat{p}_d$	$\hat{p}_d$-Carey
0.000	0.101	0.109	0.104	0.109
0.020	0.110	0.112	0.109	0.112
0.030	0.130	0.123	0.120	0.123
0.045	0.195	0.157	0.155	0.157
0.060	0.196	0.222	0.219	0.222
0.075	0.412	0.335	0.330	0.335
0.090	0.530	0.506	0.498	0.506
dose (g/kg)	$\hat{p}_{m\|\bar{d}}$-Emp	$\hat{p}_{m\|\bar{d}}$-Naive	$\hat{p}_{m\|\bar{d}}$	$\hat{p}_{m\|\bar{d}}$-Carey
0.000	0.00456	0.00190	0.00177	0.00108
0.020	0.00745	0.00910	0.00860	0.00698
0.030	0.00109	0.0198	0.0188	0.0176
0.045	0.123	0.0616	0.0593	0.0678
0.060	0.157	0.176	0.172	0.228
0.075	0.592	0.410	0.405	0.545
0.090	0.633	0.694	0.691	0.830

sponse models for death and malformation and illustrate the same phenomenon. We note that the $\hat{p}_{m|\bar{D}}(0.090g/kg)$ for Carey's model (0.830) is much greater than the empirical $\hat{p}_{m|\bar{D}}(0.090g/kg)$ (0.633) but the $\hat{p}_{m|\bar{D}}(0.075g/kg)$ for our model (0.405) and the naive model (0.410) is much lower than the empirical $\hat{p}_{m|\bar{D}}(0.075g/kg)$ (0.592), making it difficult to make a judgment on which method provides the better fit. All methods give very similar estimates for the lower doses, suggesting that the estimates for the BMDs will not vary greatly between methods.

Table 2.16 shows the BMD and BMDL estimates from each method and we indeed see that for both death models, our model has slightly higher BMDs and BMDLs. On the other hand, for the malformation model, our BMD and BMDL are higher than Carey's model. Both differences are, however, very small. Our model's BMDL is 2% smaller than Carey's BMDL for the deaths and 3% greater for the malformations. Figure 2.5 show the dose response models for death and malformation. Overall, we see that both methods give us similar results in terms of quantitative risk assessment. The relative differences also shown in Table 2.16 also indicate that the variability of the BMD is similar across methods.

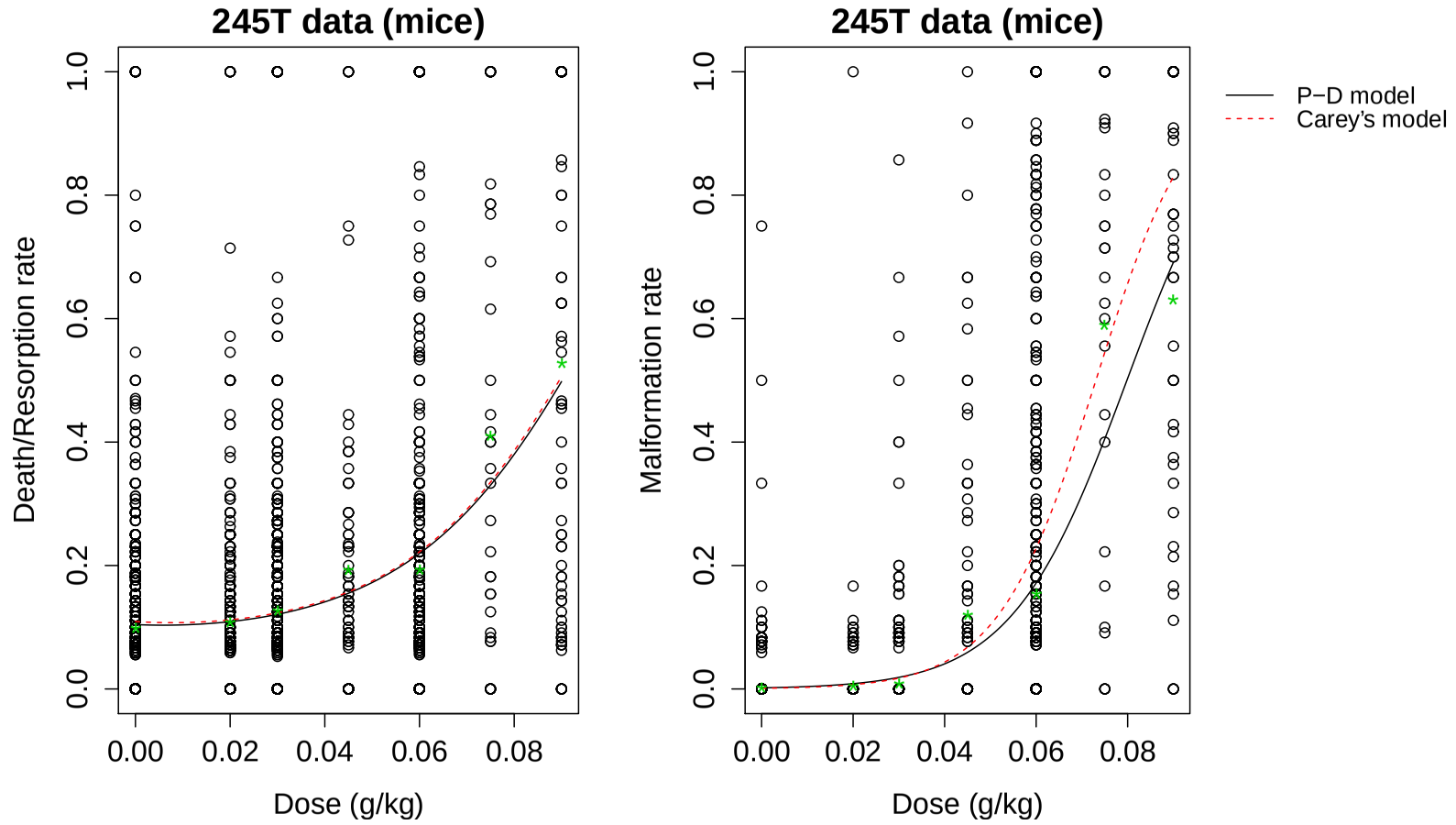

Figure 2.5: Dose Response model for death and malformation for 245T data (mice). Each dot represents the observed death/resorption rate or conditional malformation rate of a litter in the data set.

Table 2.16: Comparison of BMD and BMDL from various models

	Death model		Malformation model		
	P-D	Carey	P-D	Carey	Naive
BMD	0.0430	0.0438	0.0432	0.0417	0.0427
BMDL	0.0379	0.0387	0.0418	0.0406	0.0413
Relative Difference	0.118	0.116	0.0324	0.0264	0.0328

2.5 Discussion

In this paper, we use the Plackett-Dale distribution as a framework to develop a method for modeling hierarchical clustered developmental toxicity data. The method allows us to marginally evaluate death and conditional malformation as a function of dose while accounting for the various litter-level associations that are present in the data, including the association between death and malformation within a litter. It also allows us to model the litter-level associations as a function of dose.

One advantage of this approach is that it is more flexible in its distributional assumptions. Previously proposed methods that account for the hierarchical nature of the data, such as Christiansen's model and Carey's model, are developed from the latent normal distribution. In Christiansen's model in particular, this assumption necessitates modeling outcome-specific thresholds rather than direct probabilities, making the interpretation of dose-response parameter estimates for the probability models less intuitive than more commonly used methods for binary data, such as logistic regression. However the Plackett-Dale approach to analyzing the data allows us to choose the marginal distributions for outcomes of interest. Thus, we can choose to assume binary outcomes, death and malformation, come from a binomial distribution, and use the corresponding logistic link. Also, our method is more flexible in terms of describing the various associations among outcomes. While Christiansen's model assumes all litter-level associations can be described with one parameter, our method more flexibily estimates the litter-level associations into three different components.

The method also has the advantage of modeling the conditional malformation directly, unlike Carey's method or previous ad-hoc procedures that include an adjustment covariate to adjust for the death-malformation association. Thus, we can interpret the dose-response parameter for the conditional malformation as the overall effect of dose on malformation rate, whereas Carey's method must interpret the dose-response parameter as the effect of dose on malformation rate conditional on the adjustment covariate. It is not clear what this interpretation tells us. One possible interpretation is that it is the effect of dose on the malformation rate if there was no death-malformation correlation. Or, since the adjustment covariate is based on the residuals of the death model and thus should have mean zero, another possible interpretation is that it is the overall dose effect (since on average, the adjustment covariate should not have an effect). Even when assuming these interpretations are valid, there is still a danger in using Carey's model. Carey's adjustment covariate is derived from a latent normal framework. However, because we cannot collect data on the latent values, the adjustment covariates are transformed to the binary setting by necessity. The resulting loss of information may result in bias that may affect dose-response estimates. In contrast, our method's interpretations are much more straight forward.

We fit the model on two data sets and found probability estimates, as well as BMD and BMDLs, to be consistent with similar models at low dose levels. We also note that when we applied our method to a large data set that had strong empirical evidence suggesting that the conditional independence assumption does not hold (the 2,4,5-T study), we were able to detect a statistically significant baseline association between death outcomes and malformation outcomes. It is also worth noting the malformation model for this method does deviate somewhat from Carey's model at higher doses.

There are several avenues for further research. One is to take advantage of our estimate for ψ_{dm} to develop a method for joint risk estimation. Currently, we can estimate BMDs for each outcome, but ideally, we would like to be able to estimate joint risk so we could calculate a joint BMD. Our method estimates the relevant association, ψ_{dm}, that

ties the two outcomes together, but the nature of the Plackett-Dale distribution, in which pairs of fetuses are the unit of analysis, makes translating the information to calculate joint risk for one fetus not straightforward. Another research area would be to extend the method to include continuous outcomes such as fetal weight. Because the Plackett-Dale distribution allows flexibility in the marginal distribution, incorporating this continuous outcome should be possible, though the number of associations parameters to estimate would increase substantially.

It would also be of interest to explore the behavior of our model in different settings (for example, data with higher correlations) and compare results with Carey's model. We would be interested in whether model estimates and standard errors, as well as BMDs and BMDLs, differ substantially between the two methods. Also of interest is the small sample behavior of our method. Does our calculation method of the model parameters lead to a systematic bias of the parameter estimates? And do the theoretical standard errors for the parameters match up with what we observe in simulations? These kinds of inquiries can be made through simulations studies where we have more control over the trends in the data.

An Investigation of the Properties and Operating Characteristics of the Plackett-Dale Method for Modeling Hierarchical Outcomes in Developmental Toxicity Data via Simulations

Frederick Prichard Cudhea

Department of Biostatistics
Harvard School of Public Health

3.1 Introduction

Controlled animal studies are used to study the effects of various potentially toxic substances such as drugs or environment contaminants. In such studies, human subjects are not appropriate and researchers must rely on animals to assess toxicity from experimental data. Developmental toxicology studies are designed to examine the effect of chemical substances on developing organisms. These studies involve exposing pregnant animals (usually mice, rats, or rabbits) to a test substance during pregnancy and examining the effects on the developing implants and fetuses. Studies typically use three or four dose groups plus a control group, with at least 20 dams per dose group. The dams are sacrificed before delivery and the contents of the uterus examined. Outcomes of interest typically include number of resorptions (early deaths), number of fetal deaths, and out of the surviving fetuses: the number and type of malformations, fetal weight and fetal length. Malformations are typically categorized into three general types: Skeletal, Visceral, or External. Figure 3.1 illustrates the relationships between the various outcomes of interest (Kimmel and Price, 1990). The outcomes given the most emphasis in determining safe doses are number of embryolethalities (resorption and deaths), number of malformations, and reductions in fetal weight.

As one can see from figure 3.1, the data involve many correlations that must modeled, making statistical proper analysis challenging. For one, the major units of observation are clustered into litters so intra-litter correlation between outcomes from the same dam is expected. Secondly, among the live fetuses, there is interest in analyzing multiple outcomes (malformation status and fetal weight) from each fetus and an inter-outcome correlation is also expected. This correlation is usually not trivial and must be properly modeled for valid inference. Also, the fact that malformation status is a binary outcome while fetal weight is a continuous outcome adds another layer of complication. Third, the hierarchical relationship between the live outcomes and death further complicates interpretation the data. That is to say, the live outcomes (malformation status and fetal weight) may not only be correlated with other live fetuses, but also with dead fetuses within the same

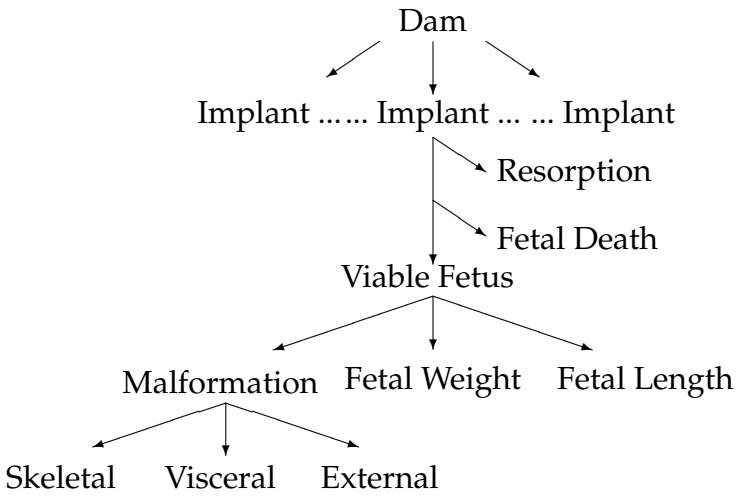

Figure 3.1: Outcomes in Developmental Toxicity

litter, and this correlation should not be ignored in the data analysis. Indeed, an examination of a large data set, from the 2,4,5-Trichlorophenoxyacetic acid developmental toxicity study (Chen and Gaylor, 1992), suggests there is a measurable and non-trivial positive association between death and conditional malformation.

3.1.1 Previous Methods

Early research focused on the problem of accounting for intra-litter correlation when only considering a single binary outcome, like death. Many important early models were developed in the late 1970's and early 1980's, including the beta-binomial model (Williams, 1975), an extension of the binomial model, and the Ochi-Prentice model (Ochi and Prentice, 1984), which used an underlying latent multivariate normal distribution to describe the intra-litter correlation. The development of generalized estimating equations (GEE) (Liang and Zeger, 1986) allowed researchers to model these data and perform accurate inference without having to correctly specify the distributions or correlations of the outcomes, making it a popular method for analyzing not just developmental toxicity data, but a wide variety of clustered discrete data.

The research on mixed outcomes has largely focused on methods based on the latent multivariate normal distribution, which gives us an intuitive and relatively simple way to characterize the correlation between malformation and fetal weight. Catalano and Ryan (Catalano and Ryan, 1992), as well as Fitzmaurice and Laird (Fitzmaurice and Laird, 1995), take advantage of the fact that the joint likelihood can be expressed as the product of the marginal distribution for weight and conditional distribution of malformation. The factorization allows weight and malformation to be modeled separately while still accounting for their correlation. Neither method is, however, conducive for formal risk estimation as joint BMDs cannot be calculated using either model. Regan and Catalano (Regan and Catalano, 1999) extend and improve on Catalano and Ryan's methodology. Their model, while still using the factorization of the latent normal distribution as a framework, allows for the estimation of the inter-outcome correlation by dose which makes joint risk, and therefore joint BMD and BMDL, possible to calculate. The correlation parameters tend to increase with dose so allowing for them to be modeled as a function of dose is an important feature of the methodology.

3.1.2 Hierarchical Relationship Between Outcomes

Less research has been done on accounting for the correlation induced by the hierarchical relationships between death and live outcomes (malformation and fetal weight). For the sake of simplicity, here we will ignore fetal weight and focus on the case where only death and malformation are outcomes of interest. Dose response modeling is interested in estimating the probability of death for a fetus as well as probability of malformation given the fetus survived and it is easy enough to estimate them separately within a dam. However, estimating joint risk of both outcomes is not as intuitive, unless we assume conditional independence. Because this assumption is not necessarily expected to be true in litter data, to compensate some models include a term (usually litter size) as a covariate for the malformation dose-response model to serve as an ad-hoc adjustment for the effect of death-rate on malformation. This approach acknowledges the hierarchical nature of the data by separating out the effect of dose and the effect of death-rate on malformation

in the modeling. However, in joint risk assessment, this hierarchical correlation is still often ignored and conditional independence is still assumed when calculating joint risk.

Most methods proposed for this issue have been inspired from previous work relying on the latent multivariate normal distribution. Christiansen (Christensen, 2004) proposes an extension to the Ochi-Prentice model, where death, malformation, and healthy outcomes are considered ordinal. Specifically, two threshold parameters, τ_m and τ_d, are used to define how the latent variable relates to the observed outcomes. Letting $\tilde{y}_{jk}$ be the latent variable for fetus j from dam k, if $\tilde{y}_{jk} < \tau_m$, then no adverse event is observed for that fetus, if $\tau_m < \tilde{y}_{jk} < \tau_d$, then a malformation is observed for the fetus and if $\tilde{y}_{jk} > \tau_d$, a fetal death is observed. The vector, $\tilde{\mathbf{y}}_k$, denoting the latent variables for the fetuses from dam k, is assumed to follow a multivariate normal distribution with mean $\mu\mathbf{1}_n$ and variance $\sigma^2((1-\rho)\mathbf{I}_n + \rho\mathbf{J}_n)$. Under these assumptions, it is possible to fit linear models for τ_m, τ_d, and ρ. Parameters μ and σ are inestimable and are assumed to be 0 and 1 respectively.

In this model, since the status of a fetus is assumed to be determined by a latent normal distribution, one correlation parameter, ρ, characterizes all three correlations of interest. Estimation under certain data conditions can be difficult. However, calculating joint risk, the risk that a fetus experiences death or malformation, is straight forward and intuitive under this model (joint risk is simply $\Phi(\tau_m)$).

Carey (Carey, 2006) develops a simpler model that still allows for conditional dependence. Essentially, she formalizes the ad-hoc approach of adding a covariate to the malformation dose-response model to adjust for the death-malformation correlation. The adjustment covariate is derived from a latent multivariate normal distribution. However, because the correlation parameters are reparameterized into the adjustment covariate, joint risk estimation is not intuitive. Also, because the adjustment variable is based on the continuous normal distribution while the observed data are binary, the actual adjustment covariate used is an approximation of the theoretical adjustment covariate. It is unclear whether these approximations are accurate or whether it potentially introduces bias.

3.1.3 Plackett-Dale

Molenbergs, Geys, and Buyse (Molenberghs et al., 2001) have taken an alternative approach, using the Plackett-Dale distribution to model malformation and fetal weight. The Packett-Dale (Plackett, 1965) (Dale, 1986) approach has an advantage over more traditional probit models in that there is flexibility in choosing the marginal distributions of the outcomes. So, for example, it is possible to assume the marginal distribution for malformation is binomial rather than what is implied, for example, by the latent normal. Let $F_{w_k}(x)$ be the cumulative distribution function for w_k, the fetal weight of a fetus from litter k and let $F_{m_k}(y)$ be the cumulative distribution function for m_k, the malformation status of a fetus from litter k. Then, if (m_k, w_k) follows a Plackett-Dale distribution, their joint cumulative distribution function is

$$F_{w_k,m_k} = \begin{cases} \frac{1+(F_{w_k}+F_{m_k})(\psi_k-1)-S(F_{w_k},F_{m_k},\psi_k)}{2(\psi_k-1)} & \psi_k \neq 1 \\ F_{w_k}F_{m_k} & \psi_k = 1 \end{cases}$$

where

$$S(F_{w_k}, F_{m_k}, \psi_k) = \sqrt{[1 + (\psi_k - 1)(F_{w_k} + F_{m_k})]^2 + 4\psi_k(1 - \psi_k)F_{w_k}F_{m_k}}$$

ψ_k, known as the global cross-ratio, defines the dependence structure of w_k and m_k,

$$\psi_k = \frac{F_{w_k,m_k}(1 - F_{w_k} - F_{m_k} + F_{w_k,m_k})}{(F_{w_k} - F_{w_k,m_k})(F_{m_k} - F_{w_k,m_k})}$$

and is used to derive the above joint cumulative density function. A psuedolikelihood based estimating equation, $pl = \sum_{k=1}^{K}\sum_{j=1}^{n_k} ln(f_{w_{jk},m_{jk}}(w, m))$, is used to estimate dose-response parameters.

Geys et al. (Geys et al., 2001) suggest, but do not implement, an extension of this method that also estimates the within-litter associations for the malformation and weight outcomes in which all associations are estimated. As we can see from figure 3.2, the within-fetus malformation-weight association is not the only source of correlation. There are also three other correlations to potentially model, but are essentially ignored in this method. For their extension, they propose defining each of these associations through a

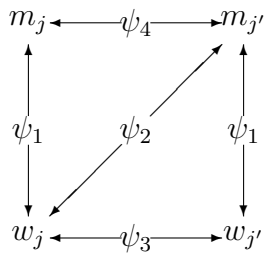

Figure 3.2: Associations present in live outcomes developmental toxicity data

global cross-ratio and then define Plackett-Dale distributions around the cross-ratios. The estimating equation are based on a log-pseudolikelihood with the following form:

$$pl = \sum_{k=1}^{K}\sum_{j=1}^{n_k} ln(f_1(w_{jk}, m_{jk})) + \sum_{k=1}^{K}\sum_{j\neq j'}^{n_k} ln(f_2(w_{jk}, m_{j'k})) + \sum_{k=1}^{K}\sum_{j'<j} ln(f_3(w_{jk}, w_{j'k})) + \sum_{k=1}^{K}\sum_{j'<j} ln(f_4(m_{jk}, m_{j'k}))$$

where f_1, f_2, f_3, f_4 are all Plackett densities characterizing different associations present in the data: ψ_1, ψ_2, ψ_3 and ψ_4.

3.1.4 Extension to Hierarchical Outcomes

Cudhea proposed a method using the Plackett-Dale framework to model dose-response for hierarchical data (Cudhea, 2013). It essentially takes the same approach of Geys et al. in their proposed extension but applies it to hierarchical data.

The various outcomes and associations of interest present within a litter can be visualized in figure 3.3. First, there is the association between two death outcomes within a cluster. For the fetuses that did not die, there is the association between two malformation outcomes within a cluster. Finally, there is the association between death outcomes and malformation outcomes, which determines how the death experience of a particular dam will affect the corresponding conditional malformation rate for the same dam.

Let us formalize the notation. Let d_{jk} be a binary random variable that is 1 if fetus j

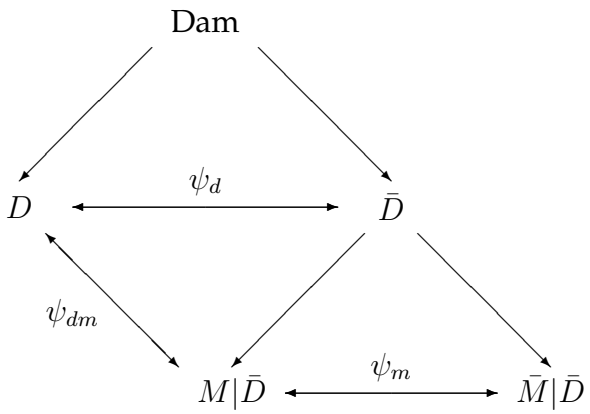

Figure 3.3: Associations present in hierarchical developmental toxicity data

from dam k is dead and 0 if alive, and let $m_{jk}|\bar{D}_{jk}$ be a binary random variable that is 1 if fetus j from dam k is malformed and 0 if not, given that fetus jk is known to not be dead. All three ψ parameters have odds ratio interpretations. Specifically:

$$\psi_d = \frac{P(D_j|D_{j'})/P(\bar{D}_j|D_{j'})}{P(D_j|\bar{D}_{j'})/P(\bar{D}_j|\bar{D}_{j'})}$$

$$\psi_m = \frac{P\left(M_j|\bar{D}_j\middle|M_{j'}|\bar{D}_{j'}\right)/P\left(\bar{M}_j|\bar{D}_j\middle|M_{j'}|D_{j'}\right)}{P\left(M_j|\bar{D}_j\middle|\bar{M}_{j'}|\bar{D}_{j'}\right)/P\left(\bar{M}_j|\bar{D}_j\middle|\bar{M}_{j'}|\bar{D}_{j'}\right)}$$

$$\psi_{dm} = \frac{P\left(M_j|\bar{D}_j\middle|D_{j'}\right)/P\left(\bar{M}_j|\bar{D}_j\middle|D_{j'}\right)}{P\left(M_j|\bar{D}_j\middle|\bar{D}_{j'}\right)/P\left(\bar{M}_j|\bar{D}_j\middle|\bar{D}_{j'}\right)}$$

where D_j is a death outcome for fetus j and M_j is a malformation outcome for fetus j. As with Regan-Catalano's method, a dose-response model can be estimated for the association parameters: ψ_d, ψ_m and ψ_{dm}.

Parameters ψ_d, ψ_m, ψ_{dm} can be thought of as global cross-ratios that define the various associations present in the data: ψ_d is the within-cluster association between death outcomes, ψ_m is the within-cluster association between malformation outcomes, and ψ_3 is the association between death outcome and malformation outcome that is induced by conditional dependence. From these cross-ratios, the joint probabilities for two deaths, two malformations (given they are not dead), and one death and one malformation (given the

malformed fetus was known not to be dead), can be derived as:

$$F_1 = P(D_j, D_{j'}) = \begin{cases} \frac{1+(2p_d)(\psi_1-1)-S(p_d,p_d,\psi_d)}{2(\psi_d-1)} & \psi_d \neq 1 \\ p_d^2 & \psi_d = 1 \end{cases}$$

$$F_2 = P(M_j|\bar{D}_j, M_{j'}|\bar{D}_{j'}) = \begin{cases} \frac{1+(2p_{m|\bar{D}})(\psi_m-1)-S(p_{m|\bar{D}},p_{m|\bar{D}},\psi_m)}{2(\psi_m-1)} & \psi_m \neq 1 \\ p_{m|\bar{D}}^2 & \psi_m = 1 \end{cases}$$

$$F_3 = P(M_j|\bar{D}_{j'}, D_{j'}) = \begin{cases} \frac{1+(p_{m|\bar{D}}+p_d)(\psi_{dm}-1)-S(p_m|\bar{D},p_d,\psi_{dm})}{2(\psi_{dm}-1)} & \psi_{dm} \neq 1 \\ p_{m|\bar{D}}p_d & \psi_{dm} = 1 \end{cases}$$

where $S(p_1, p_2, \psi) = \sqrt{[1 + (\psi - 1)(p_1 + p_2)]^2 + 4\psi(1 - \psi)p_1p_2}$.

From these joint probabilities one can derive the the probability mass functions for the paired outcomes to be:

$$G_1(d_j, d_{j'}) = \begin{cases} F_1(p_d, \psi_d) & d_j = 1, d_{j'} = 1 \\ 2(p_d - F_1(p_d, \psi_d)) & d_j \neq d_{j'} \\ 1 - 2p_d + F_1(p_d, \psi_d) & d_j = 0, d_{j'} = 0 \end{cases}$$

$$G_2(m|\bar{D}_j, m|\bar{D}_{j'}) = \begin{cases} F_2(p_{m|\bar{D}}, \psi_m) & m|\bar{D}_j = 1, m|\bar{D}_{j'} = 1 \\ 2(p_{m|\bar{D}} - F_2(p_{m|\bar{D}}, \psi)) & m|\bar{D}_j \neq m|\bar{D}_{j'} \\ 1 - 2p_{m|\bar{D}} + F_2(p_{m|\bar{D}}, \psi_m) & m|\bar{D}_j = 0, m|\bar{D}_{j'} = 0 \end{cases}$$

$$G_3(m|\bar{D}_j, d_{j'}) = \begin{cases} F_3(p_{m|\bar{D}}, p_d, \psi_{dm}) & m|\bar{D}_j = 1, d_{j'} = 1 \\ p_{m|\bar{D}} - F_3(p_{m|\bar{D}}, \psi_{dm}) & m|D_j = 1, d_{j'} = 0 \\ p_D - F_3(p_D, \psi_{dm}) & m|D_j = 0, d_{j'} = 1 \\ 1 - p_{m|\bar{D}} - p_d + F_3(p_{m|\bar{D}}, \psi_{dm}) & m|\bar{D}_j = 0, d_{j'} = 0 \end{cases}$$

The method uses a 2-step estimation procedure. The model first estimates dose response parameters for p_d and ψ_d from estimating equations based on

$$pl_1 = \sum_{k=1}^{K} \sum_{j'<j} ln(G_1(d_{jk}, d_{j'k})).$$

Then, estimates for p_m, ψ_m, and ψ_3 can then be estimated from estimating equations based on

$$pl_2 = \sum_{k=1}^{K} \sum_{j'<j} ln(G_2(m_{jk}|\bar{D}_{jk}, m_{j'k}|\bar{D}_{j'k})) + \sum_{k=1}^{K} \sum_{j\neq j'} ln(G_3(d_{jk}, m_{j'k}|\bar{D}_{j'k}))$$

by substituting parameters for p_d and ψ_d with their estimates obtained from step one.

Thus, we estimate two sets of dose-response models:

$$\eta_{k1} = \begin{pmatrix} logit(p_{d_k}) \\ log(\psi_{d_k}) \end{pmatrix} = \mathbf{X}_{k1}\boldsymbol{\beta}_1$$

$$\eta_{k2} = \begin{pmatrix} logit(p_{m|\bar{d}_k}) \\ log(\psi_{m_k}) \\ log(\psi_{dm_k}) \end{pmatrix} = \mathbf{X}_{k2}\boldsymbol{\beta}_2.$$

We use the logit-link for the probability models and the log-link for the ψ models, but other options, such as the probit-link for the probability models are also easily accommodated.

The estimating equations used to estimate $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are

$$U(\beta_1) = \sum_{k=1}^{N} \left(\frac{\partial \eta_{k1}}{\partial \boldsymbol{\beta}_1}\right)^T \left(\frac{\partial \eta_{k1}}{\partial \boldsymbol{\beta}_1}\right)^{-T} \left(\frac{\partial pl_1}{\partial \theta_{k1}}\right)$$

and

$$U(\beta_2) = \sum_{k=1}^{N} \left(\frac{\partial \eta_{k2}}{\partial \boldsymbol{\beta}_2}\right)^T \left(\frac{\partial \eta_{k2}}{\partial \boldsymbol{\beta}_2}\right)^{-T} \left(\frac{\partial pl_2}{\partial \theta_{k2}}\right)$$

respectively, where $\theta_1 = (p_d, \psi_d)$ and $\theta_2 = (p_{m|\bar{D}}, \psi_m, \psi_{dm})$.

The derivatives of the psuedolikelihoods (pl_1 and pl_2) are defined as follows:

$$\left(\frac{\partial pl_1}{\partial \theta_{k1}}\right) = \begin{pmatrix} \frac{1}{n_k - 1} \sum_{j'<j} \frac{\partial}{\partial p_{d_k}} ln(G_1(d_{jk}, d_{j'k})) \\ \\ \sum_{j'<j} \frac{\partial}{\partial \psi_{d_k}} ln(G_1(d_{jk}, d_{j'k})) \end{pmatrix}$$

and

$$\left(\frac{\partial pl_2}{\partial \theta_{k2}}\right) = \begin{pmatrix} \frac{1}{l_k - 1} \sum_{j'<j} \frac{\partial}{\partial p_{m|\bar{D}_k}} ln(G_2(m|\bar{D}_{jk}, m|\bar{D}_{j'k})) \\ + \frac{1}{n_k - 1} \sum_{j \neq j'} \frac{\partial}{\partial p_{m|\bar{d}_k}} ln(G_3(d_{jk}, m|\bar{d}_{j'k})) \\ \\ \sum_{j'<j} \frac{\partial}{\partial \psi_{m|\bar{d}_k}} ln(G_2(m|\bar{D}_{jk}, m|\bar{D}_{j'k})) \\ \\ \sum_{j \neq j'} \frac{\partial}{\partial \psi_k} ln(G_3(d_{jk}, m|\bar{D}_{j'k})) \end{pmatrix}.$$

The covariance estimates for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are

$$\hat{cov}(\hat{\boldsymbol{\beta}}_1) = \left(\sum_{k=1}^{N} \frac{\partial U_k(\boldsymbol{\beta}_1)}{\partial \boldsymbol{\beta}_1}\right)^{-1} \left(\sum_{k=1}^{N} U_k(\boldsymbol{\beta}_1) U_k(\boldsymbol{\beta}_1)^T\right) \left(\sum_{k=1}^{N} \frac{\partial U_k(\boldsymbol{\beta}_1)}{\partial \boldsymbol{\beta}_1}\right)^{-T} \Bigg|_{\boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}_1}$$

and

$$\hat{cov}(\hat{\boldsymbol{\beta}}_2) = \left(\sum_{k=1}^{N} \frac{\partial U_k(\beta_2)}{\partial \boldsymbol{\beta}_2}\right)^{-1} \left(\sum_{k=1}^{N} U_k(\boldsymbol{\beta}_2) U_k(\boldsymbol{\beta}_2)^T\right) \left(\sum_{k=1}^{N} \frac{\partial U_k(\boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}_2}\right)^{-T} \Bigg|_{\boldsymbol{\beta}_2 = \hat{\boldsymbol{\beta}}_2}$$

respectively.

The Plackett-Dale approach to analyzing developmental toxicity data has been used empirically to analyze several toxicity data sets but the method has not been systematically evaluated. Therefore, it is of great interest to study the behavior of this model in a more controlled setting. Thus, this paper sets out to investigate the properties and operating characteristics of the method via simulations. Specifically, we investigate whether our method's estimates under reasonable sample sizes are, on average, consistent with their expectations. We study this by comparing the values of parameter estimates at each dose, obtained from simulations, to their expected values at each dose, obtained by running the model on a simulated data with an extremely large sample size in order to establish "population" values.

While methods have been developed to model the issue of conditional dependence, there has yet to be a systematic comparison of these methods. In addition, there has been little research investigating how different the results from these more sophisticated models are from a naive method that assumes conditional independence. For these reasons, we also compare the behavior of the proposed method to Carey's method as well as to the naive method (which assumes conditional independence). Christiansen's method is not included in the analysis because his method does not model conditional malformation (it is also known to be computationally intensive). Because we compare our proposed method with Carey's method in this paper, we present Carey's method, as applied to our situation, when only death and malformation are outcomes of interest, more formally here. Unlike Christensen's model, Carey's likelihood uses two latent variables, one for death and one for malformation, denoted as $\tilde{d}$ and $\tilde{m}$ respectively. The two latent variables are assumed to follow a multivariate normal distribution. More specifically, for the

k-th litter:

$$\begin{pmatrix} \tilde{\mathbf{d}}_k \\ \tilde{\mathbf{m}}_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_m \end{pmatrix}, \begin{pmatrix} \Sigma_d & \Sigma_{dm} \\ \Sigma_{dm} & \Sigma_m \end{pmatrix}\right)$$

where

$$\begin{aligned}
\mu_d &= (\tilde{\alpha}_0 + \tilde{\alpha}_1 dose_k)\mathbf{1}_{n_k} \\
\mu_m &= (\tilde{\beta}_0 + \tilde{\beta}_1 dose_k)\mathbf{1}_{l_k} \\
\Sigma_d &= \sigma_d^2((1-\rho_d)\mathbf{I}_{n_k} + \rho_d \mathbf{J}_{n_k}) \\
\Sigma_m &= \sigma_m^2((1-\rho_m)\mathbf{I}_{l_k} + \rho_m \mathbf{J}_{l_k}) \\
\Sigma_{dm} &= \Sigma_{md}^T = \rho_{md}\sigma_m\sigma_d \mathbf{J}_{n_k \times l_k}
\end{aligned}$$

and l_k denotes the number of live fetuses while n_k denotes the number of implants in litter k.

Given the above likelihood, the marginal distribution of death and conditional distribution of fetal weight and malformation can be expressed as:

$$\begin{pmatrix} \tilde{\mathbf{d}}_k \\ \tilde{\mathbf{m}}_k | \tilde{\mathbf{d}}_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_{m|d} \end{pmatrix}, \begin{pmatrix} \Sigma_d & \mathbf{0}_{n_k \times l_k} & \mathbf{0}_{n_k \times l_k} \\ \mathbf{0}_{l_k \times n_k} & \mathbf{0}_{l_k} & \Sigma_{m|d} \end{pmatrix}\right)$$

where $\mu_{m|d}$ can be expressed as

$$\mu_{m|d} = (\tilde{\beta}_0 + \tilde{\beta}_1 dose) + (\rho_{md}\sigma_m)(1 + \rho_d(n_k - 1))^{-1}\left(\frac{\Sigma_{j=1}^{n_k}\tilde{d}_{ij} - n_k(\tilde{\alpha}_0 + \tilde{\alpha}_1)dose}{\sigma_d}\right).$$

Note that $\mu_{m|d}$ can be expressed as the sum of marginal model for latent malformation plus an adjustment covariate that is a function of the mean standardized residuals from the fetal death model. While the adjustment term is a bit complicated and includes parameters from the latent theory that are not estimable, this theoretical model is used to motivate a simpler, more practical adjustment term:

$$\mu_{m|d} = (\beta_0 + \beta_1 dose) + \beta_2\left(\frac{\bar{d}_k - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)}{\sqrt{\Phi(\hat{\alpha}_0 - \hat{\alpha}_1 dose)[1 - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)]/n_k}}\right).$$

Mean models are then fit using GEEs with the following dose-response framework:

$$E[d_{jk}]/\sqrt{Var(d_{jk})} = \Phi(\alpha_0 + \alpha_1 dose_k)$$
$$E[m_{jk}]/\sqrt{Var(m_{jk})} = \Phi(\beta_0 + \beta_1 dose_k)$$

To enable easy comparison between our model and Carey's model, we use a logit model version of her method rather than the proposed probit model. Given the two link functions tend to estimate similar dose-response trends in practice, we believe the derived the adjustment covariate derived by Carey will still apply when the logit link function is used. Thus, in our comparisons, the Carey models was fit using the following dose-response models:

$$\begin{aligned} logit(E[d_{jk}]) &= \alpha_0 + \alpha_1 dose_k \\ logit(E[m_{jk}]) &= \beta_0 + \beta_1 dose_k \\ &\quad + \beta_2 \left(\frac{\bar{d_k} - logit^{-1}(\hat{\alpha}_0 + \hat{\alpha}_1 dose)}{\sqrt{logit^{-1}(\hat{\alpha}_0 + \hat{\alpha}_1 dose)[1 - logit^{-1}(\hat{\alpha}_0 + \hat{\alpha}_1 dose)]/n_k}} \right) \end{aligned} \tag{3.1}$$

Finally, the paper investigates the sensitivity of dose-response parameters to the model choice for the association parameters. Since the association parameters are modeled simultaneously along with the probability parameters, how we choose to model the association parameters will affect the resulting probability models. Thus, how sensitive the probability parameters are to mis-modeling the ψ models is of great interest, and the simulations give us an opportunity to study, on average, this sensitivity.

3.2 Simulations

3.2.1 Simulation Methodology

We conducted a simulation study to investigate the behavior of our model. Because the model is not based on a single full-likelihood, it is impossible to simulate data from a true distribution consistent with our model. Furthermore, simulating from a Plackett-Dale distribution can be fairly complex. Instead, here we simulate from a multivariate

latent normal distribution, and account for conditional dependence in a manner similar to Carey's model. This allows for a simpler simulation model and facilitates more intuitive comparisons between previously proposed models, specifically Carey's model. Using a multivariate normal framework for simulating the data means that we cannot calculate theoretical ψ values, however, we can determine arbitrarily accurate estimates for the "true" ψ values, as will be shown. Basing the simulation method on Carey's method also allows for flexibility in how each of the three association parameters vary with dose. Thus, it is possible to simulate data where we know fitting a linear dose-response model for the ψ parameters is accurate.

We present the latent normal framework used in simulating the data more formally below: For litter k,

$$\begin{pmatrix} \tilde{\mathbf{d}}_k \\ \tilde{\mathbf{m}}_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_m \end{pmatrix}, \begin{pmatrix} \Sigma_d & \Sigma_{dm} \\ \Sigma_{dm} & \Sigma_m \end{pmatrix}\right)$$

where

$$\begin{aligned}
\mu_d &= (\tilde{\alpha}_0 + \tilde{\alpha}_1 dose_k)\mathbf{1}_{n_k} \\
\mu_m &= (\tilde{\eta}_0 + \tilde{\eta}_1 dose_k)\mathbf{1}_{l_k} \\
\Sigma_d &= \sigma_d^2((1-\rho_d)\mathbf{I}_{n_k} + \rho_d \mathbf{J}_{n_k}) \\
\Sigma_m &= \sigma_m^2((1-\rho_m)\mathbf{I}_{l_k} + \rho_m \mathbf{J}_{l_k}) \\
\Sigma_{dm} &= \Sigma_{md}^T = \rho_{md}\sigma_m\sigma_d \mathbf{J}_{n_k \times l_k}.
\end{aligned}$$

Here, $\tilde{\mathbf{d}}_k$ and $\tilde{\mathbf{m}}_k$ denote the latent variable vectors for death and malformation, respectively while n_k and l_k denote the number of implants and litter size for litter k. We use a factorization argument to to re-express the joint density as

$$\begin{pmatrix} \tilde{\mathbf{d}}_k \\ \tilde{\mathbf{m}}_k | \tilde{\mathbf{d}}_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_{m|d} \end{pmatrix}, \begin{pmatrix} \Sigma_d & \mathbf{0}_{n_k \times l_k} \\ \mathbf{0}_{l_k \times n_k} & \Sigma_{m|d} \end{pmatrix}\right)$$

where

$$\begin{aligned}
\mu_{m|d} &= (\beta_0 + \beta_1 dose) + (\rho_{md}\sigma_m)(1+\rho_d(n_k-1))^{-1}\left(\frac{\Sigma_{j=1}^{n_k}\tilde{d}_{ij} - n_k(\tilde{\alpha}_0 + \tilde{\alpha}_1)}{\sigma_d}\right) \\
\Sigma_{m|d} &= \sigma_m^2((1-\rho_m)\mathbf{I}_{l_k} + \rho_m \mathbf{J}_{l_k}) - \rho_{md}^2\sigma_w^2 n_k(1+\rho_d(n_k-1))^{-1}\mathbf{J}_{\mathbf{l_k}}
\end{aligned}$$

Table 3.1: Simulation Parameters by Dose

dose (g/kg/day)	c_{d_k}	c_{m_k}	ρ_d	ρ_m	ρ_{dm}
0	-1.175	-1.200	0.000	0.000	0.000
0.75	-1.075	-0.900	0.066	0.054	0.060
1.5	-0.960	-0.590	0.142	0.123	0.141
2.0	-0.725	0.110	0.300	0.300	0.300
dose (g/kg/day)	p_d	$p_{m\|\bar{D}}$	ψ_d	ψ_m	ψ_{dm}
0	0.120	0.115	0.994	1.000	1.002
0.75	0.142	0.180	1.241	1.177	1.208
1.5	0.169	0.264	1.555	1.377	1.495
3.0	0.235	0.501	2.382	1.960	2.163

and use this latent distribution to simulate our data. In practice, for each dam, we simulate the death latent outcomes from a $N(\mathbf{0}_{n_k}, \Sigma_d)$ distribution and then use a dose-specific cutoff, c_{d_k}, to determine whether a particular fetus is dead or alive (a cutoff of 0 would mean there is a 50% chance the fetus is dead). In simulating the corresponding malformation data for the same dam, we simulate from a $N(\mathbf{0}_{l_k}, \Sigma_{m|d})$ distribution and then use $c_{m_k} - (\rho_{md}\sigma_m)(1+\rho_d(n_k-1))^{-1}\left(\frac{\Sigma_{j=1}^{n_k}\tilde{d}_{ij}}{\sigma_d}\right)$ as the cutoff for that dam, where c_{m_k} is the cutoff (independent of the death outcomes for the litter) for malformation.

We simulated 5,000 datasets. Each data set had 100 dams equally distributed among four dose groups to conform to typical toxicity study sample sizes. The dose groups we used were 0 g/kg/day (control), 0.75 g/kg/day, 1.5 g/kg/day, and 3 g/kg/day. Each dam had 15 implants, again consistent with studies in rodents and rabbits. The cutoffs and correlation parameter values we used for each dose, as well as the corresponding probabilities and ψs for these parameter values are shown in Table 3.1.

The probabilities and ψ values were estimated by running an intercept model via our method on four data sets, one per dose group, each with 125,000 dams. The parameter estimates from each intercept model are considered the "true" values for our simulation study. Our goal was to eliminate as much lack-of-fit as possible so we could evaluate any potential bias of the method independent of bias due to model lack-of-fit. The parameter values were chosen so that the relationship between $logit(p)$ and dose, as well the rela-

tionship between $log(\psi)$ and dose, would be linear. This linearity was established for all five dose-response parameters via trial and error. Because the method is not based on a full-likelihood method, and because of the complexity of the Plackett-Dale distribution, it is not practical to simulate the data in such a way that the true values for each parameter in each dose-response model is a known quantity. However, in order to study the method's consistency with respect to its expectations, we need to understand the true values.

The parameter values are chosen to be consistent with the trends typically seen in toxicity studies. Specifically, with increasing dose, we would expect to see an increase in death rate, a relatively much higher increase in conditional malformation rate corresponding to a dose increase, and an increase in all three correlation parameters with dose. In particular, we emulated the trends observed in the EG mice data set.

It is somewhat more difficult to ascertain what would be an appropriate range of values for the ρ parameters. While our simulation scheme is essentially the same framework assumed for Carey's model, her model considers these parameters ancillary and are not directly estimated. To obtain an idea of appropriate values for ρ, we fit a GEE intercept model (assuming exchangeability) for each dose on the EG mice data and examined the estimates of the correlation parameters, shown in Table 3.2. We note that, while the change in association parameters was not statistically significant in this analysis, the table suggests that there is a trend for association to increase with dose. In particular, ρ_m appears to range from near 0 to around 0.37. For the sake of simplicity, we chose to range our ρ values from 0 to 0.3 for all 3 ρ parameters. This allows for a fairly strong overall trend for association with dose that allows us to observe the behavior of our model in a somewhat complex situation in terms of correlation structure. Given that a potential concern for our model is that the probability models may be susceptible to bias when the association parameters are large, we thought it prudent to look at the model's behavior with somewhat high values for the ρ parameters.

Table 3.3 gives the true values for $logit(p_d)$, $logit(p_{m|\bar{D}})$, $lm(\psi_d)$, $lm(\psi_m)$, $lm(\psi_{dm})$ by

Table 3.2: Correlation estimates by dose for EG mice data

dose (g/kg/day)	ρ_d	ρ_m
0	0.0415	-0.00557
0.75	0.0786	0.0170
1.5	0.0557	0.274
3	0.0929	0.366

Table 3.3: True values for $logit(p_d)$, $logit(p_{m|\bar{D}})$, $lm(\psi_d)$, $lm(\psi_m)$, and $lm(\psi_{dm})$ by dose

dose (g/kg/day)	$logit(p_d)$	$logit(p_{m\|\bar{D}})$	$lm(\psi_d)$	$lm(\psi_m)$	$lm(\psi_{dm})$
0	-1.99	-2.04	-0.00582	-0.000173	0.00196
0.75	-1.80	-1.52	0.216	0.163	0.189
1.5	-1.59	-1.03	0.441	0.320	0.402
3.0	-1.18	0.00415	0.868	0.671	0.771

each dose group, as calculated from running the intercept model on 125,000 simulated data sets with the parameter specifications used for our study (shown in Table 3.1). Figure 3.4 shows these points and straight line fits for each of the five parameters. From these figures we can visually verify that the goodness of fit of the linear model with the chosen link functions is very high.

While we treat these values as the true outcomes in our study, ultimately they are all estimates obtained from a very large simulated data set, and thus have corresponding confidence intervals. These confidence intervals and their lengths are shown in Table 3.2.1 for the $logit(p_d)$ and $logit(p_{m|\bar{D}})$ parameters and in Table 3.2.1 for $log(\psi_d)$, $log(\psi_m)$, and $log(\psi_{dm})$. Table 3.6 shows the lengths of confidence intervals relative to the range for each parameter, by dose (range being defined as the difference between the parameter value at the highest dose and the lowest dose, which in this case is 3β where β is the slope of the parameter's dose response model). We feel that the lengths of the confidence intervals are short enough that the sample size of 125,000 dams is sufficiently large to approximate the true expected value for our proposed model.

The estimation of the model parameters were programmed and executed in R (version 2.13). For both steps of the estimation procedure, the Newton-Raphson algorithm

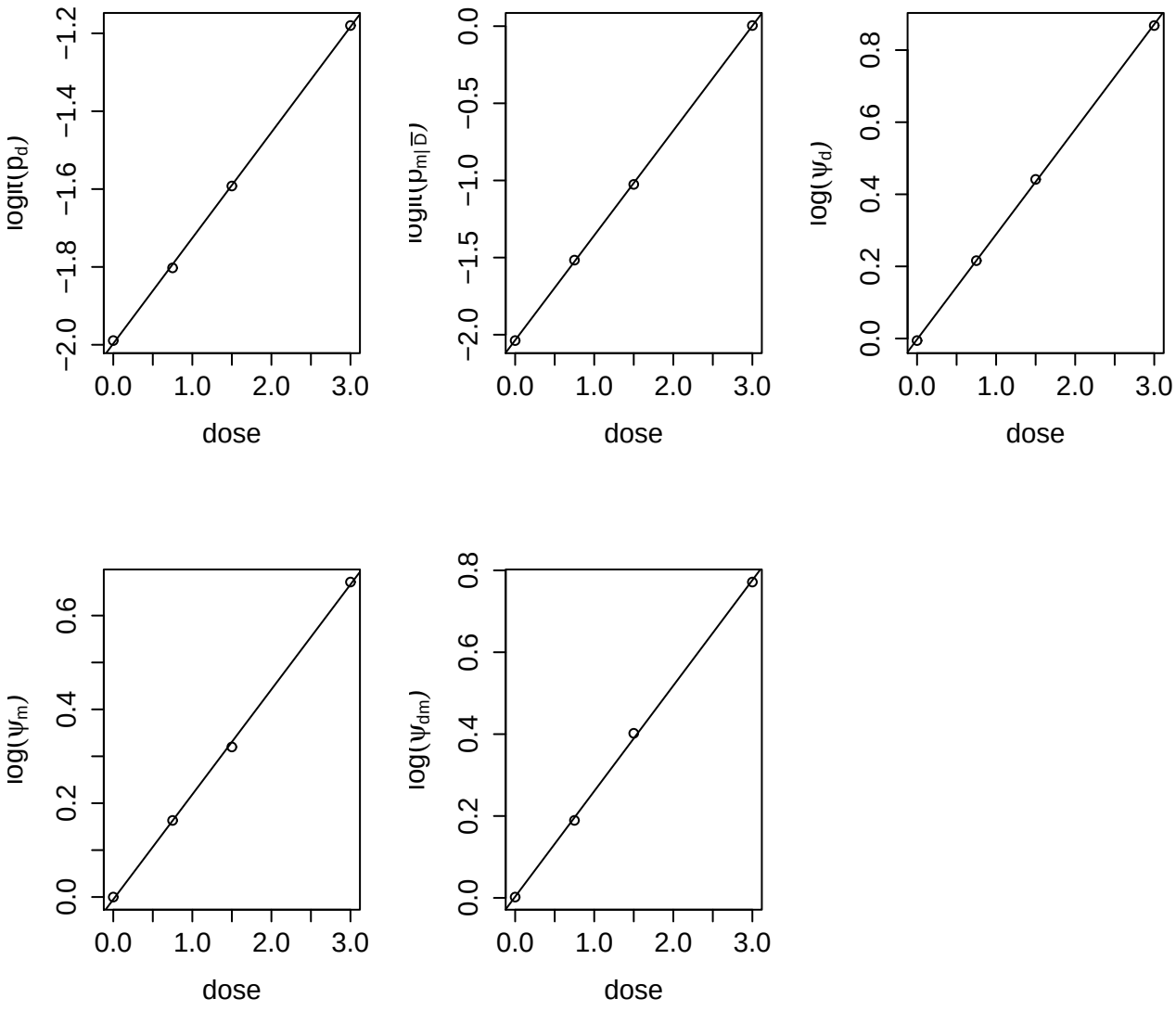

Figure 3.4: True values of $logit(p_d)$, $logit(p_{m|\bar{D}})$, $lm(\psi_d)$, $lm(\psi_m)$, and $lm(\psi_{dm})$ by dose with linear model fit

Table 3.4: Confidence intervals and their ranges for the estimated true probability parameter estimates by dose

	$logit(p_d)$		$logit(p_{m\|\bar{D}})$	
dose	CI	length	CI	length
0	(-1.994, -1.984)	0.00876	(-2.042, -2.033)	0.00955
0.75	(-1.807, -1.798)	0.00967	(-1.521, -1.512)	0.00910
1.5	(-1.597, -1.587)	0.0107	(-1.030, -1.021)	0.00917
3.0	(-1.186, -1.174)	0.0124	(-0.00112, 0.00940)	0.0105

Table 3.5: Confidence intervals and their ranges for the estimated true probability parameter estimates by dose

	$lm(\psi_d)$		$lm(\psi_m)$		$lm(\psi_{dm})$	
dose	CI	length	CI	length	CI	length
0	(-0.0108, -0.0008)	0.0100	(0.0062, 0.0059)	0.0121	(-0.0020, 0.0059)	0.0079
0.75	(0.210, 0.221)	0.0114	(0.158, 0.169)	0.0105	(0.185, 0.193)	0.0080
1.5	(0.435, 0.448)	0.0129	(0.315, 0.325)	0.0105	(0.398, 0.406)	0.0089
3.0	(0.860, 0.876)	0.0158	(0.665, 0.678)	0.0134	(0.765, 0.777)	0.012

Table 3.6: CI length relative to range for $logit(p_d)$, $logit(p_{m|\bar{D}})$, $lm(\psi_d)$, $lm(\psi_m)$, and $lm(\psi_{dm})$ by dose

dose (g/kg/day)	$logit(p_d)$	$logit(p_{m\|\bar{D}})$	$lm(\psi_d)$	$lm(\psi_m)$	$lm(\psi_{dm})$
0	0.0108	0.00467	0.0115	0.0180	0.0102
0.75	0.0119	0.00446	0.0130	0.0157	0.0104
1.5	0.0132	0.00449	0.0148	0.0157	0.0116
3.0	0.0154	0.00515	0.0180	0.0200	0.0157

for non-linear sets of equations is used. Both backtracking and the perturbation of the jacobian when not positive definite are implemented in the algorithm (Press et al., 2007). Functions to calculate first and second derivatives for G_1, G_2, and G_3 with respect to each parameter were created. Starting values for the p_d and $p_{m\bar{d}}$ models were calculated by running the equivalent logistic regression models with GEEs, ignoring hierarchical correlation. For the starting values for the ψ_d, ψ_m, and ψ_{dm} models, a starting value of 0.0001 is used. The simulations were done on a Rocks cluster with Sun Grid Engine (SGE), with each node having a processor speed of 2.66 GHz. The total running time for the simulations, which included simulating data sets and formatting them to allow analysis via the P-D method, the actual calculations of the model parameters for both the P-D method, Carey's method and the naive method, and the compilation of summary statistics, took 26.06 hours. The seed was set at 984.

Table 3.7: Simulation Results: Means

dose (g/kg/day)	p_d	$p_{m\|\bar{d}}$	ψ_d	ψ_m	ψ_{dm}
0	0.119	0.115	1.02	0.998	1.02
0.75	0.142	0.178	1.24	1.16	1.22
1.5	0.168	0.265	1.52	1.37	1.47
3	0.233	0.500	2.37	1.96	2.18

Table 3.8: Simulation Results: Bias

dose (g/kg/day)	p_d	$p_{m\|\bar{d}}$	ψ_d	ψ_m	ψ_{dm}
0	$-9.48 * 10^{-4}$	$-2.08 * 10^{-4}$	$2.43 * 10^{-2}$	$-2.30 * 10^{-3}$	$1.41 * 10^{-2}$
0.75	$2.88 * 10^{-4}$	$-2.21 * 10^{-3}$	$1.73 * 10^{-4}$	$-1.10 * 10^{-2}$	$1.23 * 10^{-2}$
1.5	$-8.62 * 10^{-4}$	$1.08 * 10^{-3}$	$-2.88 * 10^{-2}$	$-2.39 * 10^{-3}$	$-2.07 * 10^{-2}$
3	$1.26 * 10^{-3}$	$-1.97 * 10^{-4}$	$-1.05 * 10^{-2}$	$-1.18 * 10^{-3}$	$2.14 * 10^{-2}$

3.2.2 Simulation results

For each of the 5,000 data sets we fit the following model:

$$\begin{aligned} logit(p_d) &= \beta_{d_0} + \beta_{d_1} dose \\ ln(\psi_d) &= \alpha_{d_0} + \beta_{m_1} dose \\ logit(p_{m|\bar{d}}) &= \beta_{m_0} + \beta_{m_1} dose \\ ln(\psi_m) &= \alpha_{m_0} + \alpha_{m_1} dose \\ ln(\psi_{dm}) &= \alpha_{dm_0} + \alpha_{dm_1} dose. \end{aligned} \tag{3.2}$$

Out of the 5,000 datasets, the method successfully converged 4,997 times. Table 3.7 shows the means for the estimates of interest, p_d, p_m, ψ_d, ψ_m, ψ_{dm}, by dose. Tables 3.8 and 3.9 show the bias and percentage bias, respectively, for each parameter by dose (under the heading "P-D method" for Table 3.9). We see that the bias is relatively small and there does not seem to be any consistent direction or pattern for any particular dose group or parameter.

Table 3.10 shows the mean parameter estimates for the model as well as the mean and empirical standard deviations for each parameter (under the heading "P-D method"). We

Table 3.9: Simulation Results: Percentage Bias by dose for the P-D method, Carey's method, and the naive method

	P-D method					Carey's method		Naive method	
dose (g/kg/day)	p_d (%)	$p_{m\|\bar{d}}$ (%)	ψ_d (%)	ψ_m (%)	ψ_{dm} (%)	p_d (%)	$p_{m\|\bar{d}}$ (%)	p_d (%)	$p_{m\|\bar{d}}$ (%)
0	-0.79	-0.18	2.45	-0.23	1.40	-0.79	-1.98	0.79	-0.12
0.75	0.20	-1.22	0.01	-0.93	1.01	0.21	0.45	0.21	-1.32
1.5	-0.51	0.41	-1.85	-0.17	-1.38	-0.51	4.11	-0.51	0.0058
3	-0.53	-0.04	-0.44	-0.06	0.99	-0.51	5.55	-0.51	-0.76

note that the mean and empirical standard deviations are fairly close though the empirical standard deviations are slightly larger, with the exception of β_{d_0}.

We also examined the coverage of the parameter values for each dose. Typically, these are not calculated in a standard simulation analysis. Instead, the coverage and bias for the model parameters would be calculated. However, in our case, because our method is not based on a full likelihood model, we feel it is more appropriate to examine estimates for each dose, where we we have a very precise estimate for the true values. Table 3.11 shows the coverage for each probability and ψ for each dose. The coverage is consistently lower than 95% regardless of parameter for dose. This may have to do with the fact that our calculation method seems to slightly underestimate the standard deviation of the parameters. We also note that the coverage seems to decrease as dose increases. However, the coverage is fairly close to 0.95 at the lower doses.

Table 3.10: Mean, Empirical Standard Deviation, and Mean Theoretical Standard Deviation for Parameter Estimates for the P-D method, Carey's method, and the naive method

P-D method			
parameter	mean	emp sd	mean theoretical sd
$\hat{\beta}_{d_0}$	-2.00	0.138	0.135
$\hat{\beta}_{d_1}$	0.269	0.0947	0.0910
$\hat{\alpha}_{d_0}$	0.00138	0.183	0.171
$\hat{\alpha}_{d_1}$	0.276	0.126	0.115
$\hat{\beta}_{m_0}$	-2.05	0.140	0.135
$\hat{\beta}_{m_1}$	0.684	0.0884	0.0833
$\hat{\alpha}_{m_0}$	-0.0197	0.186	0.176
$\hat{\alpha}_{m_1}$	0.221	0.118	0.110
$\hat{\alpha}_{dm_0}$	-.00649	0.137	0.132
$\hat{\alpha}_{dm_1}$	0.252	0.0959	0.0930
Naive method			
parameter	mean	emp sd	mean theoretical sd
$\hat{\beta}_{d_0}$	-2.01	0.139	0.136
$\hat{\beta}_{d_1}$	0.269	0.0957	0.0932
$\hat{\beta}_{m_0}$	-2.05	0.142	0.138
$\hat{\beta}_{m_1}$	0.697	0.0898	0.0885
Carey's method			
parameter	mean	emp sd	mean theoretical sd
$\hat{\beta}_{d_0}$	-2.01	0.139	0.136
$\hat{\beta}_{d_1}$	0.269	0.0957	0.0932
$\hat{\beta}_{m_0}$	-2.09	0.145	0.139
$\hat{\beta}_{m_1}$	0.749	0.0960	0.0788
$\hat{\beta}_{m_2}$	0.333	0.0660	0.0656

Table 3.11: Coverage by dose for each parameter

dose (g/kg/day)	p_d	p_m	ψ_d	ψ_m	ψ_{dm}
0.0	0.945	0.941	0.938	0.935	0.943
0.75	0.941	0.940	0.920	0.923	0.940
1.5	0.934	0.935	0.902	0.926	0.923
3.0	0.928	0.923	0.879	0.888	0.927

3.2.3 Comparison to Carey's method and Naive method

We also fit Carey's model (3.1) on the same 5,000 simulated data sets to compare to the P-D method. Specifically, we fit the following model:

$$\begin{aligned} logit(p_d) &= \beta_{d_0} + \beta_{d_1} dose \\ logit(p_{m|\bar{D}}) &= \beta_{m_0} + \beta_{m_1} dose \\ &\quad + \beta_2 \left(\frac{\bar{d}_k - logit^{-1}(\hat{\beta}_{d_0} + \hat{\beta}_{d_1} dose)}{\sqrt{logit^{-1}(\hat{\beta}_{d_0} - \hat{\beta}_{d_1} dose)[1 - logit^{-1}(\hat{\beta_{d_0}} + \hat{\beta}_{d_1} dose)]/n_k}} \right) \end{aligned}$$

where d_k is the observed death rate for dam k and n_k is the number of implants for dam k.

The naive model is simply fitting the logistic regression model for p_d and $p_{m|\bar{D}}$ using GEE, without any consideration for correlation in the hierarchical outcomes. Specifically the models fit are:

$$\begin{aligned} logit(p_d) &= \beta_{d_0} + \beta_{d_1} dose \\ logit(p_{m|\bar{D}}) &= \beta_{m_0} + \beta_{m_1} dose. \end{aligned}$$

Table 3.10 shows the mean parameter estimates for Carey's model as well as the mean and empirical standard deviations for each parameter. As expected, we see that conditional malformation dose response is quite different between the two methods. We also observe that the empirical standard deviations seem to be slightly higher for the equivalent parameters from the P-D model. This is possibly due to lack of fit in Carey's mode and the naive model. However, given that this is true even for the death models, and both models give near identical means for their parameter estimates, it may indicate that the P-D model gives slightly more stable results, despite the added complexity. We also observe that the mean theoretical standard deviation also tends to be slightly lower than the empirical standard deviation for all three methods.

We also evaluated the bias of the three methods. Table 3.9 shows the percentage bias for p_d and $p_{m|\bar{D}}$ for each dose group. Not surprisingly, the differences in bias are small

between all methods for the death model. For the malformation model, we find that the P-D method is less biased than Carey's method, especially for the higher dose groups. This is expected, since the parameters of the simulations were chosen to fit the P-D model. The bias of the naive model is actually comparable to the P-D model. Only the percentage bias for the last two dose group is substantially different for the two methods, and even there, the percent bias does not exceed 1%.

3.2.4 Sensitivity to ψ model specification

We also examine the sensitivity of the p_d and $p_{m|\bar{d}}$ parameters to what models are fit to the ψ parameters. In the simulation studies above we fit linear models to the each of the three ψ parameters that we know to be true. In practice though, it may be difficult to detect a trend in these second order parameters because the standard deviations are much higher. Indeed, even in the 2,4,5-T study, with a sample size of over 2,500 dams, a linear trend in dose for the ψ_{dm} was not statistically significant (Cudhea, 2013). Thus, incorrectly fitting a constant model for these association parameters is entirely plausible, making it important to assess how the p_d and p_m model parameters are affected when oversimplified association model parameters are used. To examine this sensitivity, we fit two alternative models to the same 5,000 simulated data sets. One in which the ψ_{dm} parameter is assumed to be constant across doses and another in which all three ψ parameters are assumed to be constant across doses. Specifically, the two alternative models are as follows:

$$\begin{aligned} logit(p_d) &= \beta_{d_0} + \beta_{d_1} dose \\ ln(\psi_d) &= \alpha_{d_0} + \beta_{m_1} dose \\ logit(p_{m|\bar{d}}) &= \beta_{m_0} + \beta_{m_1} dose \\ ln(\psi_m) &= \alpha_{m_0} + \alpha_{m_1} dose \\ ln(\psi_{dm}) &= \alpha_{dm_0} \end{aligned} \tag{3.3}$$

Table 3.12: Mean Parameter Estimates for models (3.2), (3.3), and (3.4)

parameter	mean (model 3.2)	mean (model 3.3)	mean (model 3.4)
$\hat{\beta}_{d_0}$	-2.00	-2.00	-2.01
$\hat{\beta}_{d_1}$	0.269	0.269	0.269
$\hat{\alpha}_{d_0}$	0.00138	0.00143	0.484
$\hat{\alpha}_{d_1}$	0.276	0.276	NA
$\hat{\beta}_{m_0}$	-2.05	-2.05	-2.05
$\hat{\beta}_{m_1}$	0.684	0.681	0.681
$\hat{\alpha}_{m_0}$	-0.0197	-0.0191	0.477
$\hat{\alpha}_{m_1}$	0.221	0.221	NA
$\hat{\alpha}_{dm_0}$	-.00649	0.424	0.424
$\hat{\alpha}_{dm_1}$	0.252	NA	NA

and

$$\begin{aligned} logit(p_d) &= \beta_{d_0} + \beta_{d_1} dose \\ ln(\psi_d) &= \alpha_{d_0} \\ logit(p_{m|\bar{d}}) &= \beta_{m_0} + \beta_{m_1} dose \\ ln(\psi_m) &= \alpha_{m_0} \\ ln(\psi_{dm}) &= \alpha_{dm_0}. \end{aligned} \tag{3.4}$$

The mean parameter estimates for all three models are shown in Table 3.12. The percentage biases by dose for p_d and $p_{m\bar{D}}$ for each of the three models are shown in Table 3.13. Both tables show that the bias on p_d and $p_{m|\bar{d}}$ introduced by misfitting the ψ parameters is minimal. Oversimplifying the ψ models does seem to slightly bias the estimates for the probabilities in higher doses, but even in those cases the percent bias does not exceed 1%. Likewise, oversimplifying the ψ_{dm} appears to have little effect on the model parameters for ψ_d and ψ_m. Table 3.13 also compares the percent bias for these parameters between models 3.3 and 3.4 and the results do not seem to indicate that the ψ_d and ψ_m parameters are more sensitive to misspecification of the ψ_{dm} model than the p_d and $p_{m|\bar{D}}$ parameters are.

Table 3.13: Simulation Results: Percentage Bias of parameters for models 3.2, 3.3, and 3.4

	model 3.2		model 3.3		model 3.4	
dose (g/kg/day)	$p_d(\%)$	$p_{m\|\bar{D}}(\%)$	$p_d(\%)$	$p_{m\|\bar{D}}(\%)$	$p_d(\%)$	$p_{m\|\bar{D}}(\%)$
0	-0.79	-0.18	-0.78	0.021	-0.78	0.032
0.75	0.21	-1.22	0.21	-1.22	0.20	-1.21
1.5	-0.51	0.41	-0.51	0.27	-.51	0.26
3	-0.51	-0.04	-0.54	-0.34	-.52	0.35
dose (g/kg/day)	$\psi_d(\%)$	$\psi_m(\%)$	$\psi_d(\%)$	$\psi_m(\%)$		
0	2.45	-0.23	2.44	-0.18		
0.75	0.01	-0.93	0.01	-0.93		
1.5	-1.85	-0.17	-1.85	-0.23		
3	-0.44	-0.06	-0.44	-0.23		

3.3 Discussion

In this paper, we conduct a simulation study to explore the small sample behavior of the P-D method for model developmental toxicity data without assuming conditional independence. The method, based on the Plackett-Dale distribution, allows for the evaluation death and conditional malformation outcomes as a function of dose while accounting for the various litter-level associations that are present in the data, including the association between death outcomes and malformations within a litter. It also allows for modeling the litter-level association with an odds-ratio interpretations, including as function of dose. Theoretically, the model has some advantages over previously proposed models that also relax the conditional independence assumption. One advantage of this approach is that it is more flexible in its distributional assumptions. Since the method does not rely on the latent normal distribution, it allows for a theoretical justification to use the intuitive and widely prevalent logistic link function over the probit link function. It also not only allows for the possibility that litter-level associations can change with dose, but that different types of litter level associations can increase at different rates. It also models the conditional malformation probability directly, unlike previous methods.

The simulation study showed the method's variance estimators are close to the empirical variances and that the estimation procedure did appear be able to estimate the

parameters with minimal bias. Thus, the small sample behavior of the model seems to be consistent with its large sample expectations. We also compared the simulations results from the P-D model to Carey's model and the naive model and found that for the comparable parameters (the parameters for the death model and the intercept parameter for the malformation model), the estimates are fairly similar for the naive and the P-D models. We also found that using the equivalent parameter for Carey's model to estimate conditional tended to overestimate the conditional malformation probabilities. This is not surprising since Carey's model estimates the conditional malformation probability slope given the adjustment covariate which depends on the death rate of the litter. It is worth noting the simulation study indicates that if one only uses β_{m_1} to describe the increase in conditional malformation probabiilty by dose, then one will overestimate the trend. We also observed that, despite the complexity of the P-D model, the variance estimates for the parameters were fairly close on average as well (for the comparable parameters).

We also assessed the how sensitive the probability model parameters were to misfitting the ψ model parameter and found that oversimplifying the ψ models had only a small impact on the estimates for p_d and $p_{m\bar{D}}$. This is not surprising given how similar the probability estimates are between the P-D model and the naive model.

There are several avenues for further research to explore. One is to extend the method to include continuous fetal weight as an outcome of interest. Because the Plackett-Dale distribution allows flexibility in the marginal distribution, incorporating this continues outcome should be possible (though the number of associations parameters to estimate would increase substantially). It would also be of interest to explore the behavior of our model in different settings (for example, data with higher correlations) and compare results with Carey's model. In particular, It would be of interest to see whether model estimates and standard deviations, differ substantially between the two methods.

The ultimate goal of the data analysis, however, is to fit a dose-response model to each outcome, and to use these models to inform safe doses for regulation purposes. A key step translating the dose-response model to a 'safe' a dose is the calculation of the

BMD (benchmark dose) and BMDL (the associated lower bound) (Gaylor et al., 1998), a process referred to as quantitative risk estimation, part of the larger and more general goal of quantitative risk assessment. The BMD is defined as the dose that corresponds to a given percent increase in risk above background (usually 5 or 10%). The BMDL is the statistical lower-bound (usually 95%) of the BMD, and is the quantity most useful in assessing and establishing safety standards. Often, a BMDL is calculated for each outcome and the smallest is chosen, which can lead to underestimating the safe dose and ignores any correlation. A more valid approach would be to calculate a joint BMD that accounts for the combined risk of all outcomes. This approach requires that joint risk, the probability of any adverse outcome, be estimable, meaning that a joint distribution for the outcomes must be specified and that relevant inter-outcome correlations must be estimated. For methods where this is not possible (often because inter-outcome correlations are not estimated), conditional independence is assumed. That is, it is assumed that the live outcomes (malformation and fetal weight) are independent of the death outcomes. In other words, the death rate of a litter does not inform the malformation rate (or fetal weights) of the litter. Thus, for example, if we are only interested in death and malformation outcomes, the joint risk, $P(Adverse\ Event) = P(Dead\ or\ Malformed)$, simplifies to $1 - (1 - P(Dead) * (1 - P(Malformed \mid Not\ Dead)))$. The approach, while commonly used, is not satisfying, as there is no theoretical basis for this assumption.

Therefore, it is of great interest to take advantage of our estimate for ψ_{dm} to develop a method for joint risk estimation. Currently, we can estimate BMDs separately for each outcome, but ideally, we would like to be able to estimate joint risk so we could calculate a joint BMD. Our method estimates the relevant association, ψ_{dm}, that ties the two outcomes together, but the nature of the Plackett-Dale distribution, in which pairs of fetuses are the unit of analysis, makes translating the information to calculate joint risk for one fetus not straightforward. Thus, pursuing a method to calculate joint BMDs from the P-D model, as well as Carey's method, would be worthwhile. By applying these methods to real data, as well as various simulated scenarios, we can compare the various methods to one another.

Methods for BMD and BMDL Estimation for Outcomes in Developmental Toxicity Studies

Frederick Prichard Cudhea

Department of Biostatistics
Harvard School of Public Health

4.1 Introduction

Controlled animal studies are used to study the effects of various potentially toxic substances such as pharmaceuticals or environmental contaminants. In such studies, human subjects are often not appropriate and researchers must rely on animal studies to assess toxicity from experimental data. Developmental toxicology studies are designed to examine the effect of chemical substances on developing organisms. These studies involve exposing pregnant animals (usually mice, rats, or rabbits) to a test substance during pregnancy and examining the effects on the fetuses. Studies typically use three or four dose groups plus a control group, with at least 20 dams per dose group. The dams are sacrificed before delivery and the contents of the uterus examined. Outcomes of interest typically include number of resorptions (early deaths), number of fetal deaths, and out of the surviving fetuses, the number and type of malformations, fetal weights and fetal lengths. Malformations are typically categorized into three general types: Skeletal, Visceral, or External. Figure 4.1 illustrates the relationships between all the various outcomes of interest (Kimmel and Price, 1990). The outcomes given the most emphasis in determining safe doses are number of embryolethalities (resorption and deaths), number of malformations, and reductions in fetal weight.

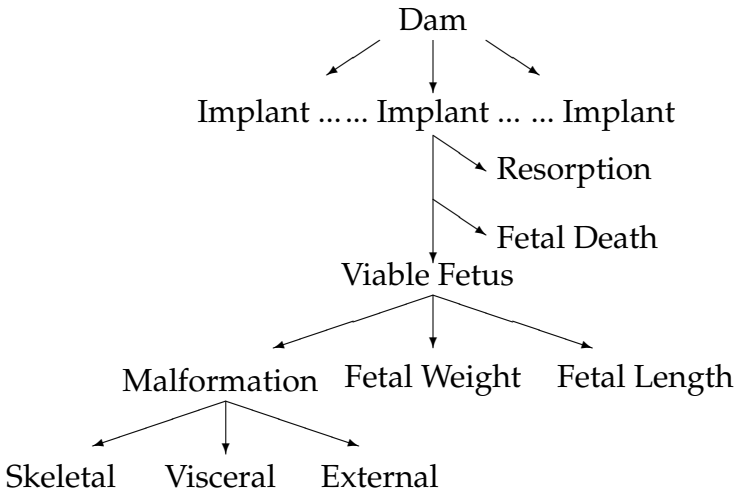

Figure 4.1: Outcomes in Developmental Toxicity

As one can see from Figure 4.1, the data involve many possible correlations that must be modeled, making proper analysis challenging. The major units of observation are clustered into litters so intra-litter correlation between outcomes from the same dam is expected. Secondly, among the live fetuses, there are multiple outcomes (malformation status and fetal weight) from each fetus and an inter-outcome correlation is also expected. This correlation is usually not trivial and must be properly modeled for valid statistical inference. The fact that malformation status is a binary outcome while fetal weight is a continuous outcome adds another layer of complication. Third, the hierarchical relationship between the live outcomes and death further complicates interpretation the data. That is to say, the live outcomes (malformation status and fetal weight) may not only be correlated with other live fetuses, but also with dead fetuses within the same litter, and this correlation cannot be ignored in the data analysis.

The ultimate goal of the data analysis is to measure dose-response relationships in each outcome, and to use these models to inform safe doses for regulation purposes. This process, referred to as quantitative risk estimation, is part of the larger goal of quantitative risk assessment. In the past, a key step in translating the dose-response model to a 'safe' a dose was the calculation of the NOAEL (no-observed-adverse-effect level), the highest observed dose in which the chemical has a statistically significant effect. However, this metric has major flaws, in that they are restricted to actual doses from the experiment, they encourage poor (small sample size) study designs, and does not provide a corresponding estimate of associated risk (Crump, 1984). Thus this statistic has been replaced in favor of the more precise and dose-response model driven BMD (benchmark dose) and BMDL (benchmark dose - lower bound) (Gaylor et al., 1998).

The BMD is defined as the dose that corresponds to a given percent increase in risk above background (usually 5 or 10 %). The increase is known as the benchmark response, or BMR. The BMD is obtained from solving $\frac{p(dose)-p(0)}{1-p(0)} = BMR$.

The BMDL is the statistical lower-bound (usually 95%) of the BMD, and is the quantity most useful in assessing and establishing safety standards. There are several meth-

ods to calculate the BMDL including the standard Wald approach, where $BMDL_{.95} = BMD - 1.645 * sqrt(var(BMD))$, and the maximum likelihood approach, where the $BMDL_{.95}$ is the dose that satisfies $2(l_{max} - l_1) = 1.645^2$ and minimizes the BMD (l_{max} is the unrestricted maximized log-likelihood and l_1 is the log-likelihood under some constraint). The Wald approach is known to yield unstable results and the maximum likelihood approach requires a model that assumes a full likelihood distribution. For this paper, we use a method for calculating a BMDL proposed by Kimmel and Gaylor (Kimmel and Gaylor, 1988). In this method, we calculate the dose that corresponds to the specified excess risk for the 95% upper confidence bound of the dose-response curve. In practice, this means the $BMDL_{.95}$ is the dose that solves $\hat{r}(dose) + 1.645se(\hat{r}(dose)) = BMR$ where $se(\hat{r}(dose)$ is calculated via the delta method.

Characterizing risk using the methods described above works well when only considering a single adverse outcome. However, in many cases, more than one outcome is of interest (for example death and malformation). The simplistic approach to determining a safe dose in this scenario is to calculate a BMD for each outcome and then choose the lowest one. This approach does not take into account the joint toxic effects of the outcomes, however, and can lead to an underestimation of the safe dose, especially when the outcomes have low correlation (Ryan, 1992). A better approach is to calculate a BMD based on joint risk that combines all outcomes of interest. In the context of developmental toxicology, where death and malformation are outcomes of interest, this means calculating one BMD based on $P(M \cup D|dose)$ rather than choosing the smaller of two BMDs based on $P(D|dose)$ and $P(M|\bar{D}, dose)$. This approach requires that joint risk, the probability of any adverse outcome, be estimable, meaning that a joint distribution for the outcomes must be specified and that relevant inter-outcome correlations must be estimated. For methods where this is not possible (often because inter-outcome correlations are not estimated), conditional independence is assumed. That is, it is assumed that the live outcomes (malformation and fetal weight) are independent of the death outcomes. In other words, the death rate of a litter does not inform the malformation rate of the litter. Thus, the joint

risk, $P(Adverse\ Event) = P(Dead\ or\ Malformed)$, simplifies to

$$P(Adverse\ Event) = 1 - (1 - P(Dead) * (1 - P(Malformed \mid Not\ Dead))). \quad (4.1)$$

The approach, while commonly used, is not satisfying, as there is no theoretical basis for this assumption and it ignores potentially substantial correlation in the litter.

In this paper, we present three methods to analyze such data, and then propose three approaches to calculate the joint risk BMD that take advantage of the unique properties of the method. We then evaluate and compare the resulting joint BMDs in real data as well as in simulated scenarios.

4.2 Methods

4.2.1 Naive Method

The naive method simply assumes conditional independence, and thus ignores the hierarchical correlation present in the data. In other words, death outcomes and malformation outcomes (conditional on the fetuses being alive) are modeled separately. The intra-litter correlations, between death outcomes and between malformation outcomes, need not be ignored and are accounted for via using GEEs (Liang and Zeger, 1986) to estimate the model parameters and their standard errors.

4.2.2 Carey's Method

Carey (Carey, 2006) develops a straightforward model that allows for conditional dependence. The method, taking a similar approach to Regan's model for fetal malformation and fetal weight (Regan and Catalano, 1999), essentially formalizes the ad-hoc approach of adding an adjustment covariate to the malformation dose-response model to adjust for the death-malformation correlation.

Carey's likelihood uses two latent variables, one for death and one for malformation, denoted as $\tilde{d}$ and $\tilde{m}$ respectively. The two latent variables are assumed to follow a multivariate normal distribution. More specifically, for the k-th litter:

$$\begin{pmatrix} \tilde{\mathbf{d}}_k \\ \tilde{\mathbf{m}}_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_m \end{pmatrix}, \begin{pmatrix} \Sigma_d & \Sigma_{dm} \\ \Sigma_{dm} & \Sigma_m \end{pmatrix} \right) \tag{4.2}$$

where

$$\begin{aligned} \mu_d &= (\tilde{\alpha}_0 + \tilde{\alpha}_1 dose_k)\mathbf{1}_{n_k} \\ \mu_m &= (\tilde{\beta}_0 + \tilde{\beta}_1 dose_k)\mathbf{1}_{l_k} \\ \Sigma_d &= \sigma_d^2((1-\rho_d)\mathbf{I}_{n_k} + \rho_d \mathbf{J}_{n_k}) \\ \Sigma_m &= \sigma_m^2((1-\rho_m)\mathbf{I}_{l_k} + \rho_m \mathbf{J}_{l_k}) \\ \Sigma_{dm} &= \Sigma_{md}^T = \rho_{md}\sigma_m\sigma_d \mathbf{J}_{n_k \times l_k} \end{aligned} \tag{4.3}$$

and l_k denotes the number of live fetuses while n_k denotes the number of implants in litter k.

Given the above likelihood, the marginal distribution of death and conditional distribution of fetal malformation can be expressed as:

$$\begin{pmatrix} \tilde{\mathbf{d}}_k \\ \tilde{\mathbf{m}}_k | \tilde{\mathbf{d}}_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_{m|d} \end{pmatrix}, \begin{pmatrix} \Sigma_d & \mathbf{0}_{n_k \times l_k} & \mathbf{0}_{n_k \times l_k} \\ \mathbf{0}_{l_k \times n_k} & \mathbf{0}_{l_k} & \Sigma_{m|d} \end{pmatrix} \right)$$

where $\mu_{m|d}$ is given by

$$\mu_{m|d} = (\tilde{\beta}_0 + \tilde{\beta}_1 dose) + (\rho_{md}\sigma_m)(1 + \rho_d(n_k - 1))^{-1} \left(\frac{\sum_{j=1}^{n_k} \tilde{d}_{ij} - n_k(\tilde{\alpha}_0 + \tilde{\alpha}_1)dose}{\sigma_d} \right)$$

or the sum of the marginal model for latent malformation plus an adjustment covariate that is a function of the mean standardized residual for fetal death. While the adjustment term is a bit complicated and includes parameters from the latent theory that are not estimable, this theoretical model is used to motivate a simpler adjustment term:

$$\mu_{m|d} = (\beta_0 + \beta_1 dose) + \beta_2 \left(\frac{\bar{d}_k - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)}{\sqrt{\Phi(\hat{\alpha}_0 - \hat{\alpha}_1 dose)[1 - \Phi(\hat{\alpha}_0 + \hat{\alpha}_1 dose)]/n_k}} \right)$$

Mean models are then fit using GEEs within the following dose-response framework:

$$E[d_{jk}]/\sqrt{Var(d_{jk})} = \Phi(\alpha_0 + \alpha_1 dose_k)$$
$$E[m_{jk}]/\sqrt{Var(m_{jk})} = \Phi(\beta_0 + \beta_1 dose_k)$$

To enable easy comparison between our model and Carey's model, we use a logit model version of her method rather than the proposed probit model. Given the two link functions tend to estimate similar trends in practice, we believe the adjustment covariate derived by Carey will still apply in principle even under the logit link:

$$logit(E[d_{jk}]) = \alpha_0 + \alpha_1 dose_k$$
$$logit(E[m_{jk}]) = \beta_0 + \beta_1 dose_k + \beta_2 \left(\frac{\bar{d}_k - logit^{-1}(\hat{\alpha}_0 + \hat{\alpha}_1 dose)}{\sqrt{logit^{-1}(\hat{\alpha}_0 - \hat{\alpha}_1 dose)[1 - logit^{-1}(\hat{\alpha}_0 + \hat{\alpha}_1 dose)]/n_k}} \right)$$

Both dose-response models are fit using GEEs.

4.2.3 Plackett-Dale framework

Cudhea proposed a method using the Plackett-Dale framework to model dose-response for hierarchical data (Cudhea, 2013). It takes a similar approach to Geys (Geys et al., 2001) but applies it to hierarchical data.

The various outcomes and associations of interest present within a litter can be visualized in Figure 4.2. First, there is the association between two death outcomes within a litter. For fetuses that did not die, there is the association between two malformation outcomes within a cluster. Finally, there is the association between death outcomes and malformation outcomes, which determines how the death experience of a particular dam will affect the corresponding conditional malformation within the same dam.

Let us formalize the notation. Let d_{jk} be a binary random variable that is 1 if fetus j from dam k is dead and 0 if alive, and let $m_{jk}|\bar{D}_{jk}$ be a binary random variable that is 1 if fetus j from dam k is malformed and 0 if not, given that fetus jk is known to not be dead.

Parameters ψ_d, ψ_m, ψ_{dm} from Figure 4.2 can be thought of as global cross-ratios that

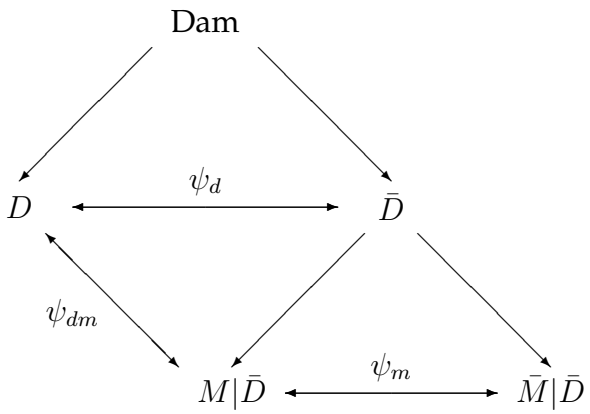

Figure 4.2: Associations present in hierarchical developmental toxicity data

define the various associations present in the data: ψ_d is the within-cluster association between death outcomes, ψ_m is the within-cluster association between malformation outcomes, and ψ_3 is the association between the death outcome and malformation outcome that is induced by conditional dependence. From these cross-ratios, the joint probabilities for two deaths, two malformations (given they are not dead), and one death and one malformation (given the malformed fetus was known not to be dead), can be derived as:

$$F_1 = P(D_j, D_{j'}) = \begin{cases} \frac{1+(2p_d)(\psi_1-1)-S(p_d,p_d,\psi_d)}{2(\psi_d-1)} & \psi_d \neq 1 \\ p_d^2 & \psi_d = 1 \end{cases}$$

$$F_2 = P(M_j|\bar{D}_j, M_{j'}|\bar{D}_{j'}) = \begin{cases} \frac{1+(2p_{m|\bar{D}})(\psi_m-1)-S(p_{m|\bar{D}},p_{m|\bar{D}},\psi_m)}{2(\psi_m-1)} & \psi_m \neq 1 \\ p_{m|\bar{D}}^2 & \psi_m = 1 \end{cases}$$

$$F_3 = P(M_j|\bar{D}_{j'}, D_{j'}) = \begin{cases} \frac{1+(p_{m|\bar{D}}+p_d)(\psi_{dm}-1)-S(p_m|\bar{D},p_d,\psi_{dm})}{2(\psi_{dm}-1)} & \psi_{dm} \neq 1 \\ p_{m|\bar{D}}p_d & \psi_{dm} = 1 \end{cases}$$

where D_j is a death outcome and M_j is a malformation outcome for fetus j, and $S(p_1, p_2, \psi) = \sqrt{[1 + (\psi - 1)(p_1 + p_2)]^2 + 4\psi(1 - \psi)p_1p_2}$.

From these joint probabilities one can derive the probability mass functions for the

paired outcomes:

$$G_1(d_j, d_{j'}) = \begin{cases} F_1(p_d, \psi_d) & d_j = 1, d_{j'} = 1 \\ 2(p_d - F_1(p_d, \psi_d)) & d_j \neq d_{j'} \\ 1 - 2p_d + F_1(p_d, \psi_d) & d_j = 0, d_{j'} = 0 \end{cases}$$

$$G_2(m|\bar{D}_j, m|\bar{D}_{j'}) = \begin{cases} F_2(p_{m|\bar{D}}, \psi_m) & m|\bar{D}_j = 1, m|\bar{D}_{j'} = 1 \\ 2(p_{m|\bar{D}} - F_2(p_{m|\bar{D}}, \psi)) & m|\bar{D}_j \neq m|\bar{D}_{j'} \\ 1 - 2p_{m|\bar{D}} + F_2(p_{m|\bar{D}}, \psi_m) & m|\bar{D}_j = 0, m|\bar{D}_{j'} = 0 \end{cases}$$

$$G_3(m|\bar{D}_j, d_{j'}) = \begin{cases} F_3(p_{m|\bar{D}}, p_d, \psi_{dm}) & m|\bar{D}_j = 1, d_{j'} = 1 \\ p_{m|\bar{D}} - F_3(p_{m|\bar{D}}, \psi_{dm}) & m|D_j = 1, d_{j'} = 0 \\ p_D - F_3(p_D, \psi_{dm}) & m|D_j = 0, d_{j'} = 1 \\ 1 - p_{m|\bar{D}} - p_d + F_3(p_{m|\bar{D}}, \psi_{dm}) & m|\bar{D}_j = 0, d_{j'} = 0 \end{cases}$$

The method uses a 2-step estimation procedure, The model first estimates dose response parameters for p_d and ψ_d, and then uses the parameter estimate for p_d to estimate p_m, ψ_m and ψ_{dm}. The exact estimation equations used for both steps are described by Cudhea (Cudhea, 2013), as are the formulas for the covariance estimates.

4.3 Estimation of Joint Risk

The methods presented that don't assume conditional independence divide the correlation parameter into three separate association parameters (between death outcomes, between malformation outcomes, and between death and malformation outcomes) and are based on distributions that explicitly connect hierarchical association (between death and malformation outcomes) and joint risk. However, no intuitive formula for $P(M \cup D)$ exists for these two models. This is so because, for Carey's model, the association parameters are not directly estimated, and for the P-D model, a full likelihood distribution is not assumed or used in the estimation of the parameters. These two methods do, however, estimate parameters that measure the level of correlation between the death and malformation outcomes. The challenge then, is to incorporate this information into joint risk for death and malformation so that a joint BMD can be calculated.

4.3.1 Naive Method

The naive method for calculating joint risk is to use formula (4.1). The formula is derived from the assumption that conditional dependence is true. In other words, it assumes that the death rate of a litter will not inform the conditional malformation rate that same litter, the very assumption that we seek to relax in our models. Thus, it is a somewhat unsatisfying to use this formula for BMD and BMDL estimation. Nevertheless, it is a simple and popular way to calculate joint risk, so the option is explored in this paper. Theoretically, ignoring the hierarchical correlation in joint risk, as this method does, should lead to overestimating joint risk and thus underestimating the joint BMD as long as the hierarchical association is positive at all doses. Thus, in all practical scenarios, we expect using this naive joint risk formula to be a conservative method for calculating a joint BMD.

4.3.2 Mean adjustment method

In principle, the mean adjustment method is not very different from the naive method. It uses the same joint risk formula, but it uses an alternative way to calculate $P(M|\bar{D})$. Recall that, in Carey's method, conditional malformation is modeled as a function of dose and also an adjustment covariate that is a function of litter size and death rate. If we choose to model $logit(p_{m|\bar{D}})$ linearly with dose then we would fit the model $logit(p_{m|\bar{D}}) = \beta_0 + \beta_1 dose_k + \beta_2 \left(\frac{\bar{d}_k - logit(\hat{\alpha}_0 + \hat{\alpha}_1 dose)}{\sqrt{logit(\hat{\alpha}_0 - \hat{\alpha}_1 dose)[1 - logit(\hat{\alpha_0} + \hat{\alpha}_1 dose)]/n_k}} \right)$. Theoretically, on average, the adjustment covariate should be zero, since both $\bar{d}_k$ and $logit(\hat{\alpha}_0 + \hat{\alpha}_1 dose)$ are unbiased estimators of the death rate at the specified dose. Therefore, in the naive method, β_2 is ignored in calculating $P(M|\bar{D})$. An alternative way of calculating joint risk is to include the adjustment covariate in the calculation of the conditional malformation probability. Since the adjustment covariate is dam specific rather than dose specific, (because n_k and $\bar{d}_k$ varies by dam), we propose using the observed mean adjustment covariate for the purpose of calculating joint risk. By including β_2 into the formula for joint risk, we can incorporate information about the association between death and malformation outcomes into the calculation of the BMD and BMDL. Theoretically, the mean of the

adjustment covariate is expected to be zero, and thus, the estimate of the BMD should not differ much from using the naive method. In practice however, the mean adjustment covariate is often non-zero (examples can be found in appendix A.2). Thus, it is possible that ignoring the adjustment covariate will lead to a biased estimate of $P_{m|\bar{D}}$ and therefore, of the BMD as well. Furthermore, the inclusion of the β_2 parameter in the calculation of the BMDL will actually take into account the uncertainty associated with that parameter. Using the naive method with Carey's model ignores this uncertainty and thus possibly underestimates the BMDL.

4.3.3 Placket-Dale method

Recall that in the Plackett-Dale model, the parameter that measures the association between death and malformation is ψ_{dm}. Thus, to incorporate the association between hierarchical outcomes in a joint risk formula for the P-D model must entail including this parameter. We propose using the formula

$$P(Adverse\ Event) = 1 - G_3(0,0).$$

Recall that $G_3(m|\bar{D}_j, d_{j'})$ is the full probability mass function derived from $F_3(m|\bar{D}_j, d_{j'})$, the joint probability that fetus j is malformed and fetus j' is dead, given that fetus j is not dead. Parameter ψ_{dm} characterizes the association between fetus j and fetus j'. Thus, $G_3(0,0)$ is the probability, for a given pair of fetuses from the same dam, that neither are malformed or dead, given that one is known to be not dead. It is not the probability that a single fetus experiences no adverse outcomes. Although somewhat ad-hoc, we believe that G_3 does have properties that make it a plausible candidate to use as a probability mass function for the outcome of one fetus. First, in the absence of any correlation between death and live outcomes in a dam, $G_3(0,0)$ simplifies to $P_d(0)P_m(0|\bar{D})$, the probability of a fetus being healthy when conditional independence is assumed. Second, the manner in which the estimate for ψ_{dm} affects joint risk is, for the most part, intuitive. Let us assume the death rate and conditional malformation rate are the same between two dams, but one has a higher ψ_{dm}. Then, we would expect the litter with the higher ψ_{dm}

to have a higher joint risk, since an increase in ψ_{dm} results in an increase in $G_3(0,0)$ and thus, a decrease in joint risk. This is consistent with our intuitive understanding of how hierarchical outcomes correlations should affect joint risk. The higher this correlation, the higher one would expect the death outcome and the malformation outcomes to be consistent, and thus the outcome of no adverse event (no death or malformation) should be more likely.

4.4 Example

To illustrate the methods described above, we apply them to two different datasets, an NTP study examining the effects of Ethanol Glycol (EG) in mice (Price et al., 1985) and a large sample study examining the effects of 2,4,5-Trichlorophenoxyacetic Acid (2,4,5-T) in mice (Chen and Gaylor, 1992). The EG data set is an example of a fairly typical study, with 98 dams and four dose groups. The 2,4,5-T data set is much larger, with 2455 dams and 7 dose groups. The following five methods are used to calculate the BMD and BMDL for both datasets:

1. Using the naive model with the naive joint risk formula
2. Using the P-D model with the naive joint risk formula
3. Using Carey's model with the naive joint risk formula
4. Using the P-D model with the P-D formula for joint risk
5. Using Carey's model with the mean-adjustment joint risk.

The naive formula is applicable for all models and thus examined for all three models discussed here. The model fits used are the same as those used by Cudhea (Cudhea, 2013), and are shown in appendix A.1.

Table 4.1: BMD, BMDL, and Relative Difference (RD) for EG Mice data

Method	1	2	3	4	5
BMD	0.504	0.512	0.503	0.517	0.503
BMDL	0.420	0.434	0.407	0.389	0.424
RD	0.168	0.153	0.191	0.248	0.158

4.4.1 NTP Study of EG in Mice

Table 4.1 shows the BMD and BMDL estimates for five different methods for the EG study. The BMD estimates range from 0.503 g/kg to 0.517 g/kg. We see that using the P-D model, regardless of what method we use to calculate the BMD, gives us a higher BMD than the other methods. The BMDL estimates range from 0.389 to 0.434. We do not note any obvious pattern with regards to how the BMDLs differ by method. The relative differences between BMD and BMDL (defined to be $(BMD-BMDL)/BMD$) range from 0.153 to 0.248. The P-D joint risk formula includes the parameter ψ_{dm}, a second order parameter that tends to have estimates with a high variance, and that added uncertainty in the joint risk estimates is expected to be reflected in a higher variance BMD estimate. Thus, it is not surprising that the relative difference between BMD and BMDL when using method 4 (P-D model with P-D risk formula) is much greater than the relative difference using method 2 (P-D model with the naive joint risk formula). It is, however, somewhat surprising that we observe the relative difference of method 5 (Carey's model with the joint risk formula that includes the adjustment term) is actually smaller than that of method 3 (Carey's model with the naive joint risk formula) since method 5 includes an additional parameter in its joint risk formula. This is especially interesting since we do not observe a strong positive correlation between the adjustment term and the other parameter estimates in the conditional malformation model. The P-D model assumes ψ_{dm} is constant across dose and estimates ψ_{dm} for this data set to be a relatively low 1.24 (95% CI of (0.912, 1.70)) which explains the homogeneous BMD estimates from the five different estimation methods.

Table 4.2: BMD, BMDL, and Relative Difference (RD) for 2,4,5-T Mice data

Method	1	2	3	4	5
BMD	3.45	3.41	3.46	3.46	3.47
BMDL	3.09	3.06	3.12	2.92	3.13
RD	0.104	0.103	0.097	0.158	0.098

4.4.2 Study of 2,4,5-T in Mice (CD-1 strain)

Table 4.2 shows the risk estimates for the same five methods for a study of 2,4,5-T on mice (CD-1 strain). All five methods give similar estimates for the BMDs, ranging from 3.41 dg/g to 3.47 dg/g. The Plackett-Dale joint risk formula gives the lowest BMDL while Carey's model (regardless of what which joint risk formula is used) give the largest BMDLs. The P-D-model assumes ψ_{dm} is constant across dose and estimates ψ_{dm} to be 1.85 (95% CI of (1.60, 2.13)), much higher than what was estimated for the EG study, making the homogeneity BMD estimates from the five different estimation methods observed for this data set somewhat surprising.

4.5 Simulations

4.5.1 Methodologic development

A simulation study was conducted to examine the behavior of the five BMD and BMDL methods, under 8 different scenarios, each defined by three parameters, the increase in the magnitude of the death dose response, the magnitude of the conditional malformation dose response, and the increase in magnitude of the within-cluster correlations (including the correlation between death and malformation correlation) by dose. We evaluate the simulation design parameters in a binary fashion ("high" and "low"). Thus the eight different scenarios can be described as:

1. high p_d slope, high $p_{m|\bar{D}}$ slope, high ψ slopes

2. high p_d slope, high $p_{m|\bar{D}}$ slope, low ψs slopes
3. high p_d slope, low $p_{m|\bar{D}}$ slope, high ψs slopes
4. high p_d slope, low $p_{m|\bar{D}}$ slope, low ψs slopes
5. low p_d slope, high $p_{m|\bar{D}}$ slope, high ψs slopes
6. low p_d slope, high $p_{m|\bar{D}}$ slope, low ψs slopes
7. low p_d slope, low $p_{m|\bar{D}}$ slope, high ψs slopes
8. low p_d slope, low $p_{m|\bar{D}}$ slope, low ψs slopes

The method of simulation is based on Carey's model. The latent normal framework used is shown in equations (4.2) and (4.3).

We use a factorization argument to re-express the joint density as

$$\begin{pmatrix} \tilde{\mathbf{d}}_k \\ \tilde{\mathbf{m}}_k|\tilde{\mathbf{d}}_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_{m|d} \end{pmatrix}, \begin{pmatrix} \Sigma_d & \mathbf{0}_{n_k \times l_k} \\ \mathbf{0}_{l_k \times n_k} & \Sigma_{m|d} \end{pmatrix}\right)$$

where

$$\mu_{m|d} = (\beta_0 + \beta_1 dose) + (\rho_{md}\sigma_m)(1 + \rho_d(n_k - 1))^{-1}\left(\frac{\Sigma_{j=1}^{n_k}\tilde{d}_{ij} - n_k(\tilde{\alpha}_0 + \tilde{\alpha}_1)}{\sigma_d}\right)$$

$$\Sigma_{m|d} = \sigma_m^2((1-\rho_m)\mathbf{I}_{l_k} + \rho_m\mathbf{J}_{l_k}) - \rho_{md}^2\sigma_w^2 n_k(1 + \rho_d(n_k - 1))^{-1}\mathbf{J}_{\mathbf{l_k}}$$

and use this latent distribution to simulate the data. In practice, for each dam, we simulate the death latent outcomes from a $N(\mathbf{0}_{n_k}, \Sigma_d)$ distribution and then use a dose-specific cutoff, c_{d_k}, to determine whether a particular fetus is dead or alive (a cutoff of 0 would mean there is a 50% chance the fetus is dead). In simulating the corresponding malformation data for the same dam, we simulate from a $N(\mathbf{0}_{l_k}, \Sigma_{m|d})$ distribution and then use $c_{m_k} - (\rho_{md}\sigma_m)(1+\rho_d(n_k-1))^{-1}\left(\frac{\Sigma_{j=1}^{n_k}\tilde{d}_{ij}}{\sigma_d}\right)$ as the cutoff for that dam, where c_{m_k} is the cutoff (independent of the death outcomes for the litter) for malformation.

For each scenario, 5,000 data sets were simulated, each with 4 dose groups (0, 0.75, 1.5, and 3.0), 25 dams per dose group, and 15 fetuses per dam. The cutoff values and

correlation values by dose for each simulation scenario can be found in appendix A.3. In some simulations scenarios, not all data sets were successfully simulated. However, these cases compromised less than 1% of attempts.

For each scenario we modeled the data in three different ways: via the naive conditional independence method, the P-D method, and the mean-adjustment method. Specifically, for the naive method, we fit the model:

$$\begin{aligned} p_d &= logit^{-1}(\beta_{d_0} + \beta_{d_1} dose) \\ p_{m|\bar{D}} &= logit^{-1}(\beta_{m_0} + \beta_{m_1} dose) \end{aligned}$$

For the P-D method, we fit the model:

$$\begin{aligned} p_d &= logit^{-1}(\beta_{d_0} + \beta_{d_1} dose) \\ p_{m|\bar{D}} &= logit^{-1}(\beta_{m_0} + \beta_{m_1} dose) \\ \psi_d &= exp(\alpha_{d_0} + \alpha_{d_1} dose) \\ \psi_m &= exp(\alpha_{m_0} + \alpha_{m_1} dose) \\ \psi_{dm} &= exp(\alpha_{dm_0} + \alpha_{dm_1} dose) \end{aligned} \tag{4.4}$$

For Carey's method, we fit the model:

$$\begin{aligned} p_d &= logit^{-1}(\beta_{d_0} + \beta_{d_1} dose) \\ p_{m|\bar{D}} &= logit^{-1}(\beta_{m_0} + \beta_{m_1} dose + \beta_{m_2} adjustment) \end{aligned}$$

In other words, for each parameter, we fit a simple linear model with with no polynomial terms.

For some data sets, an attempt to calculate a BMD failed. In some cases, it is possible that a BMD does not exist for the fitted model due to a shallow death rate or the conditional malformation rate (or both) (or due to the event rates actually decreasing rather than increasing with dose). In these cases, which are much more common in scenarios where the death rate or conditional malformation rate are low, a joint BMD will not exist. Table 4.5.1 shows the number of times we were not able to calculate a joint BMD for each

Table 4.3: Number of failed BMD calculation attempts by scenario and BMD calculation method

	method 1	method 2	method 3	method 4	method 5
Scenario 1	0	0	0	0	0
Scenario 2	0	0	0	0	0
Scenario 3	0	0	0	0	0
Scenario 4	0	0	0	0	0
Scenario 5	0	0	0	0	0
Scenario 6	0	0	0	0	0
Scenario 7	31	34	31	105	32
Scenario 8	16	14	16	48	16

joint BMD calculation method and simulation scenario. As is evident in the table, this only occurred in scenarios 7 and 8, where both the increase in death rate by dose, and the increase in malformation rate by dose are relatively low by design. The fact that we observe more failures in scenario 7, where the correlations increase more quickly with dose, is possibly an artifact of the conditional malformation rates tending to decrease as the correlation parameters increase. It is worth noting that we do observe that method 4, which uses the P-D BMD calculation method, is more susceptible to failure than the others. This is not surprising since, in extreme circumstances, it is possible for the P-D joint risk formula to decrease with dose even as the death rate and conditional rate are estimated to increase with dose (See appendix A.5 for detail). However, given how rarely these failures occur (2.1% of the time for Scenario 7) and are only observed in extreme scenarios designed specifically for understanding performance of the methods under sub-optimal conditions, it does not appear that this is a significant practical weakness of the method.

4.5.2 Results

Figure 4.3 shows the BMD distributions for the five methods via modified boxplots for simulation scenarios 1-4. Likewise, Figure 4.4 shows the same for scenarios 5-8. These boxplots show the median BMDL (red) and empirical BMDL (green) for each method and their corresponding confidence intervals. The median BMDL is the median of the

BMDL distribution while the empirical BMDL is the 5th percentile BMDs calculated from the relevant empirical distribution. The mean BMD is also included, indicated with a blue dot. The rest of the plot is the same as a typical boxplot where the box spans the 25th percentile to the 75th percentile and the central bar signifies the median.

For all eight scenarios, we observe consistent patterns between the five methods. First, Carey's model using the naive joint risk formula always gives the lowest median BMD. On the opposite end of the spectrum, the BMDs using the P-D model with the P-D joint risk formula are consistently the least conservative. We also note that BMDs from method 1 (naive method with naive joint risk) have a very similar distribution to method 2 (P-D method with P-D joint risk), but with method 2 consistently having a slightly higher median BMD. Method 5 (mean-adjustment model with mean-adjustment joint risk) does not hold to a consistent pattern in relation to the other four methods. In the simulation scenarios where the dose-response for p_d is high (1-4), method 5 gives the second lowest median BMD, below method 1 (naive model with naive joint risk) but above method 2 (P-D model w/ P-D joint risk). In the other four scenarios, where the dose response for p_d is low, method 5 actually provides median BMDs that are higher than the median BMDs for method 1. In the case of scenarios 5 and 6 (both scenarios in which p_d dose response is low and p_m dose response is high), method 5's median BMDs are actually higher than those of method 2.

The distribution of empirical BMDLs appears to follow those for the BMDs. For scenarios 7 and 8, the empirical BMDLs have very wide confidence intervals, such that all five overlap, and thus it is difficult to discern specific patterns. The median calculated BMDLs, on the other hand, do not follow the same patterns as the median BMDs. In fact, in most situations, the confidence intervals for the calculated median BMDLs do not overlap with the confidence intervals for the empirical BMDLs. For the methods that use the naive joint risk formula, the calculated median BMDs are consistently higher than the empirical BMD. For method 4 (PD model w/ P-D joint risk), the opposite is true; the median calculated BMDL is lower than the empirical BMD. For method 5 (mean-adjustment

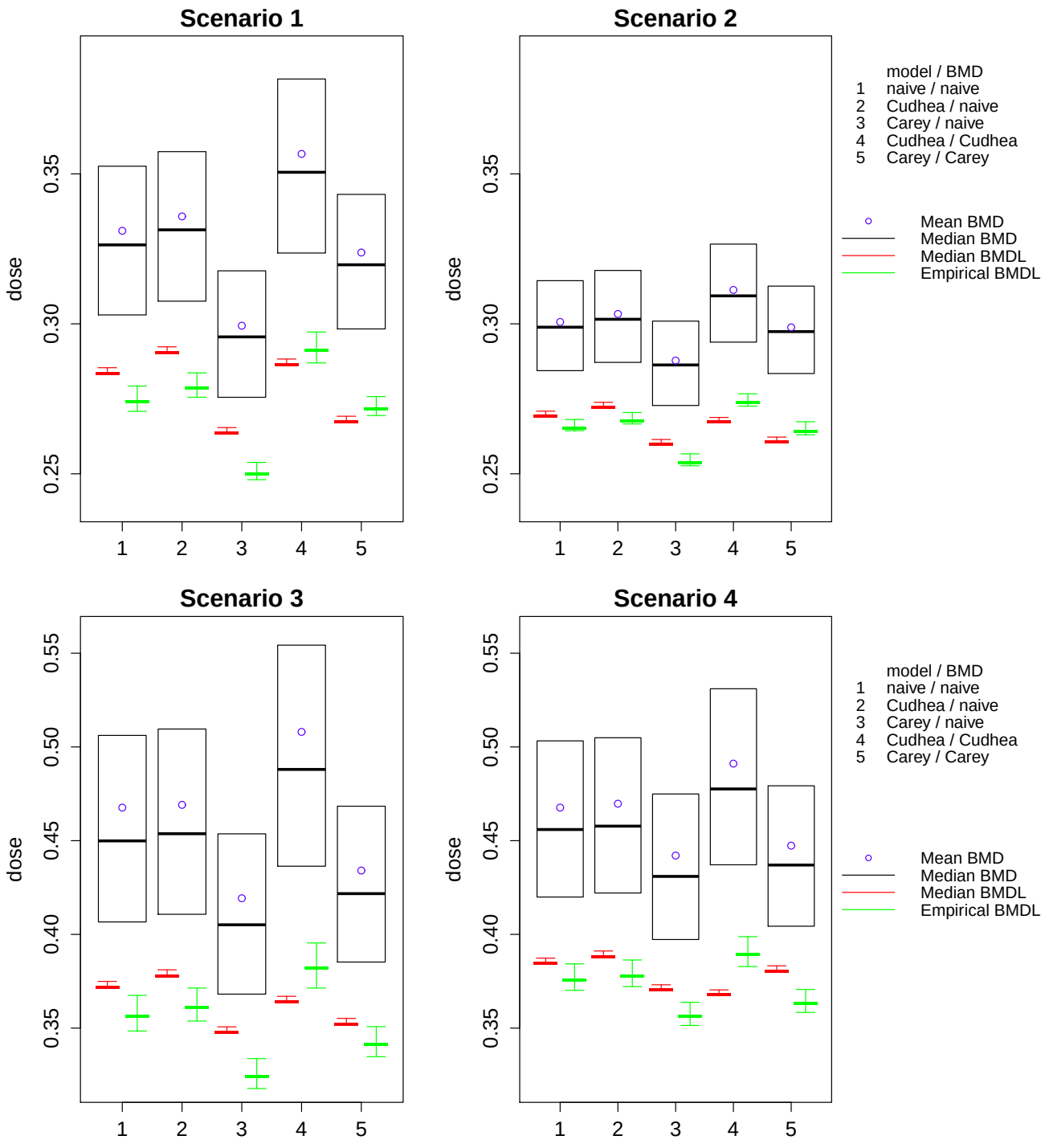

Figure 4.3: Boxplots of joint BMDs for simulation scenarios 1, 2, 3, and 4

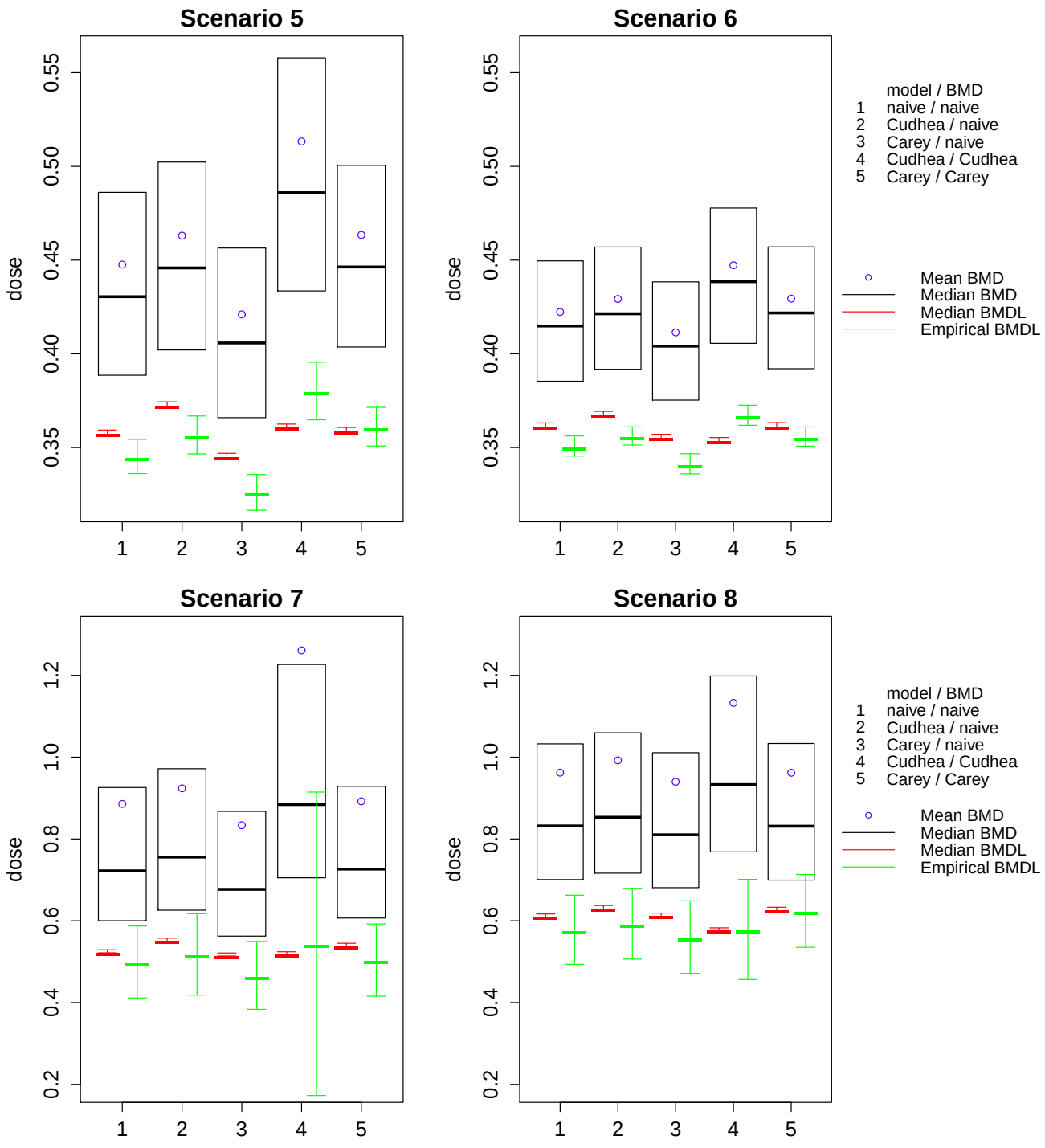

Figure 4.4: Boxplots of joint BMDs for simulation scenarios 5, 6, 7, and 8

method w/ mean-adjustment joint risk), there is no clear pattern. In scenarios 1 and 2, the median BMDL is lower than the lower bound of the empirical BMDL, in scenarios 3 and 4, the median BMDL is higher than the upper bound of the empirical BMD, while in scenarios 5 and 6, the median BMDL is contained within the confidence interval of the 95% confidence interval. Because the calculated BMDLs have a different pattern than the BMDs, determining which method is most conservative or least conservative according to BMDL is dependent on the dose-response patterns. For scenarios 1, 2, 3, 5, and 7, we observe that method 3 is still the most conservative. However, in scenarios 4, 6, and 8, it appears method 4 is most conservative. However, we do note that for scenarios 4, 6, and 8, the the difference in median BMDL for methods 3 and 4 are fairly slight whereas they are much more pronounced in methods 1, 2, 3, and 5.

Numerical mean and median summaries for the joint BMDs and BMDLs, as well as associated standard deviations, are shown in Table 4.5.2 for all 8 scenarios.

It is also worth noting that for each method and scenario, the distribution of the BMDs appear to be right-skewed. This is not surprising given the nature of the data (theoretical minimum of 0 with no theoretical maximum). In addition, the skew seems to increase as the respective median increases.

4.5.3 Relationship between BMD estimates and association parameters

We note that, as expected, the differences between the five methods seem to be more extreme in the scenarios where the dose-response for the association parameters is higher. When we compare BMD distributions from scenarios where the dose-response between high vs. low ψ values, it is clear that the difference between methods 4 and 5 (methods that account for death-malformation association in the calculation of the BMD) and methods 2 and 3 (methods that use the same modeling technique but use the naive joint risk to calculate joint BMDs) is much greater under the high ψ scenario. However, it is important to note that these scenarios don't necessarily share the same dose-response for p_d and

Figure 4.5: Mean, median and standard deviations for the joint BMD and BMDLs, as well as the empirical BMDL values, from all eight simulation scenarios for all five methods

		BMD			BMDL			
Scenario	Method	mean	median	stdev	empirical	mean	median	stdev
1	1	0.331	0.326	0.0411	0.275	0.287	0.284	0.0301
	2	0.336	0.331	0.0406	0.280	0.294	0.291	0.0306
	3	0.299	0.296	0.0346	0.251	0.267	0.265	0.0280
	4	0.357	0.351	0.0476	0.292	0.289	0.287	0.0289
	5	0.324	0.320	0.0366	0.273	0.272	0.268	0.0303
2	1	0.301	0.299	0.0229	0.266	0.271	0.270	0.0195
	2	0.303	0.302	0.0232	0.269	0.274	0.273	0.0197
	3	0.288	0.286	0.0217	0.255	0.262	0.261	0.0191
	4	0.311	0.309	0.0248	0.275	0.269	0.268	0.0188
	5	0.299	0.297	0.0223	0.265	0.263	0.262	0.0196
3	1	0.466	0.450	0.0871	0.358	0.380	0.373	0.0524
	2	0.469	0.454	0.0851	0.363	0.386	0.379	0.0526
	3	0.419	0.405	0.0762	0.326	0.355	0.349	0.0491
	4	0.508	0.488	0.105	0.383	0.371	0.365	0.0465
	5	0.434	0.421	0.0724	0.343	0.360	0.353	0.0515
4	1	0.468	0.456	0.0699	0.377	0.392	0.390	0.0457
	2	0.470	0.458	0.0699	0.379	0.396	0.390	0.0457
	3	0.442	0.431	0.0443	0.358	0.378	0.372	0.0443
	4	0.447	0.437	0.0638	0.365	0.382	0.376	0.0449
	5	0.491	0.478	0.0788	0.391	0.374	0.369	0.0391
5	1	0.448	0.430	0.0911	0.345	0.365	0.430	0.0509
	2	0.463	0.446	0.0937	0.357	0.381	0.373	0.0537
	3	0.421	0.406	0.0837	0.326	0.353	0.345	0.0523
	4	0.513	0.486	0.132	0.380	0.368	0.361	0.0479
	5	0.463	0.446	0.132	0.361	0.368	0.359	0.0586
6	1	0.422	0.415	0.0538	0.351	0.365	0.361	0.0376
	2	0.429	0.421	0.0549	0.356	0.372	0.368	0.0384
	3	0.411	0.404	0.0530	0.341	0.360	0.356	0.0385
	4	0.447	0.438	0.0612	0.367	0.357	0.354	0.0337
	5	0.429	0.422	0.0549	0.356	0.366	0.362	0.0405
7	1	0.886	0.722	0.653	0.499	0.556	0.525	0.135
	2	0.924	0.756	0.664	0.518	0.587	0.553	0.143
	3	0.834	0.677	0.518	0.466	0.553	0.518	0.148
	4	1.26	0.884	1.24	0.565	0.544	0.521	0.107
	5	0.892	0.726	0.652	0.504	0.577	0.541	0.156
8	1	0.962	0.832	0.550	0.578	0.643	0.613	0.143
	2	0.992	0.853	0.581	0.593	0.663	0.633	0.149
	3	0.940	0.810	0.542	0.560	0.646	0.615	0.150
	4	1.13	0.933	0.782	0.624	0.598	0.579	0.107
	5	0.962	0.831	0.546	0.579	0.659	0.629	0.153

$p_{m|\bar{D}}$. Indeed, we note that the BMDs from the high ψ scenarios from methods 1, 2, and 3 (all using the naive joint risk formula) are also higher than their counterparts from the low ψ scenarios. Thus, there is a possibility that the discrepency between the difference between BMDs from calculation methods that account for conditional dependence and calculation methods that ignore conditional dependence for high ψ vs low ψ scenarios is actually not related to the magnitutde of the death-malformation association and more to do with the magnitude of the risk of death and malformation outcomes.

To investigate this discrepancy while accounting for BMD magnitude, we examined a standardized version of the difference between methods 2 and 4, and the difference between methods 3 and 5. Specifically, for each simulation scenario, we report (median BMD for method 4 - median BMD for method 2) / median BMD for method 2 and (median BMD for method 5 - median BMD for method 3) / median BMD for method 3. Table 4.5.3 shows these values for each scenario. While we do observe that high ψ scenarios have a higher relative difference than their low ψ counterparts across both methods, we also note that these relative differences vary by p_d and $p_{m|\bar{D}}$ specification and still seem to depend on individual outcome risk. Therefore, individual outcome risk is still potentially a confounder (summary statistics for individual outcome risk BMDs estimates can be found in appendix A.6).

A more helpful plot might be Figure 4.6 which shows all eight relative differences for both methods against their respective naive BMDs. These plots illustrate that, for the P-D method, even when controlling for individual outcome risk, the relative differences for the BMDs tend to be higher for high ψ scenarios. While we cannot make any definitive conclusions based on these results due to small sample size, the observed trend does seem to suggest that BMD calculation methods that account for the hierarchical correlation are indeed sensitive to the hierarchical correlation. It is also worth noting that it appears that the relative difference in the P-D method seems to increase as the naive BMD increases, but no such strong trend is observed for the mean-adjustment method.

To more thoroughly understand how the differences between methods change as the

Table 4.4: Relative difference between median BMDs of method 2 and 4, and of method 3 and 5

	P-D method (method 2 vs. 4)		mean adjustment method (method 3 vs. 5)	
	high ψs	low ψs	high ψs	low ψs
high p_d, high $p_{m\|\bar{D}}$	0.0579	0.0259	0.0813	0.0389
high p_d, low $p_{m\|\bar{D}}$	0.0756	0.0433	0.0410	0.0141
low p_d, high $p_{m\|\bar{D}}$	0.0900	0.0406	0.0999	0.0438
low p_d, low $p_{m\|\bar{D}}$	0.1698	0.0936	0.0733	0.0260

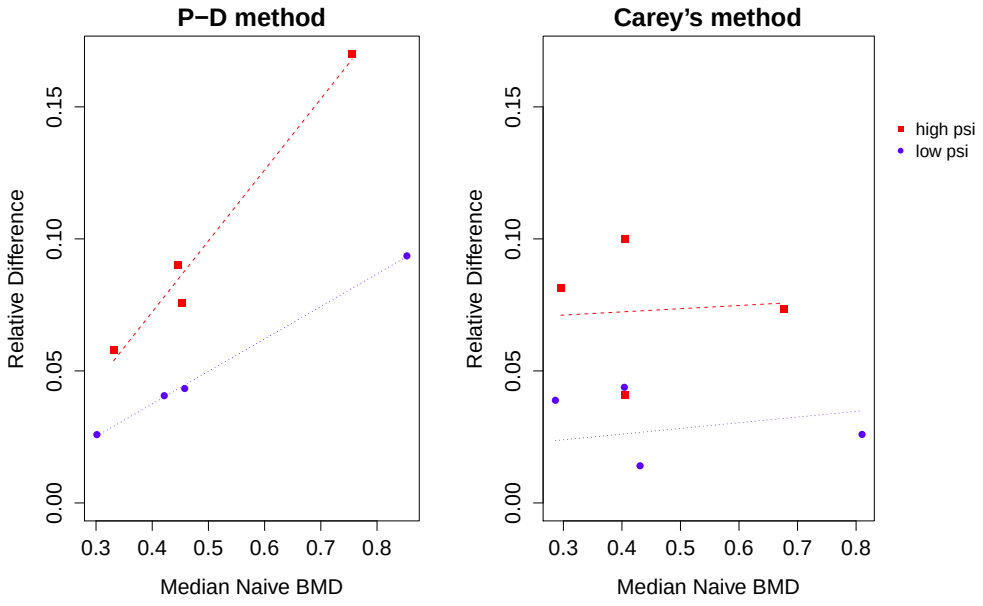

Figure 4.6: Relative Difference vs. median naive BMD for the P-D method and Carey's method

magnitude of the correlation increases, we conduct another simulation study, looking at 10 simulation scenarios in total, but only where the ρs change in the simulation parameters. The correlation parameters used for each scenario can be found in Table A.3 in appendix A.3. To distinguish these simulation scenarios from the eight used in section 4.5.2, they are labeled with an asterisk (*). For c_{d_k} and c_{m_k}, the same values are used as in scenarios 1 and 2, the ones corresponding to the high p_d, high $p_{m|\bar{D}}$ scenarios. The reason for using these cutoffs is to minimize the number of BMD calculations that fail due to the simulated study having a negative dose-response trend for one of the outcomes. The likelihood of this happening increases when combining low dose-response trends for the outcomes and high dose-response trends for the correlation parameters. Since we look at fairly high correlation patterns for this examination, using a high dose-response for the outcome probabilities seemed especially prudent in order to minimize bias resulting from failed BMD calculations.

Because these various scenarios won't necessarily have the same death and malformation probabilities for each dose, we examine the median relative differences between methods to control for individual level risk. Again, we define the relative difference between the BMD from method A from method B to be $(BMD_A - BMD_B)/BMD_B$. Figure 4.7 plots these relative differences against the median estimated α_{dm_1}, the dose-response parameter for the ψ_{dm} model, for a given simulation scenario, to evaluate how the median relative differences change as the level of hierarchical association changes. The figure shows the trend for most of the 10 relative differences examined is fairly linear or exponential. Thus, the degree to which these methods differ from each other, for the most part, appears to increase as the associations increase, and does so at a predictable rate. There are, however, two notable exceptions. The median relative difference between method 2 (naive/P-D) vs method 1 (naive/naive), as well as between method 4 (P-D / P-D) vs. method 1 (naive/naive), increase as hierarchical correlation initially increases, but then begin to decrease for scenarios where hierarchical association slopes are fairly high. For the median relative difference between methods 2 and 1, this shift occurs somewhere between scenarios 7* and 8* while for the median relative difference between methods

4 and 1, the shift occurs somewhere between scenarios 9* and 10*. The median relative difference between methods 2 and 1 actually changes from positive to negative, meaning the joint BMD estimates for method 2 are smaller for method 1 in these extreme scenarios. Since these two methods use the same naive joint risk formula, and since the death model parameter estimates have been observed to be fairly consistent for any given estimation method, it is resonable to suspect that what is driving this shift in relative difference are the parameter estimates for the malformation model. Specifically, in scenarios with extremely high correlation dose-response trends, the P-D model predicts a lower dose-response for malformation than the naive model does. It should be stressed though that scenarios 7* through 10* are not likely to be observed in practice. Thus, we do not believe this somewhat odd behavior of the P-D model based methods is a weakness in practice.

4.5.4 BMDs for individual outcomes

Figures 4.8 and 4.9 show the distribution of the individual death and malformation BMDs in comparison to the distribution of the joint BMDs via boxplots (without whiskers, as in Figures 4.3 and 4.4). As expected, the joint BMDs are consistently lower than either of the individual BMDs, once again highlighting the danger of underestimating risk when only using single-outcome BMDs for risk assessment, especially in scenarios when the risk of one outcome is not dominant relative to the other. Also as expected, the death BMDs for each method have very similar distributions, since it has been observed in previous simulations that the death model from the P-D method gives estimates very similar to simply using GEEs (Cudhea, 2013). The malformation BMDs, however, differ quite significantly based on method. The figures illustrate that the individual malformation BMDs, not the death BMDs, are a primary driver in the differences in joint BMDs between methods. In some cases, such as scenarios 3 and 4, where the death rate is overwhelmingly low compared to the malformation rate, we see that the joint BMD is not much lower than the death BMD. However, in more realistic scenarios where the conditional malformation rate is consistently higher than the death rate, the difference between joint and individual BMDs seems to be non-trivial.

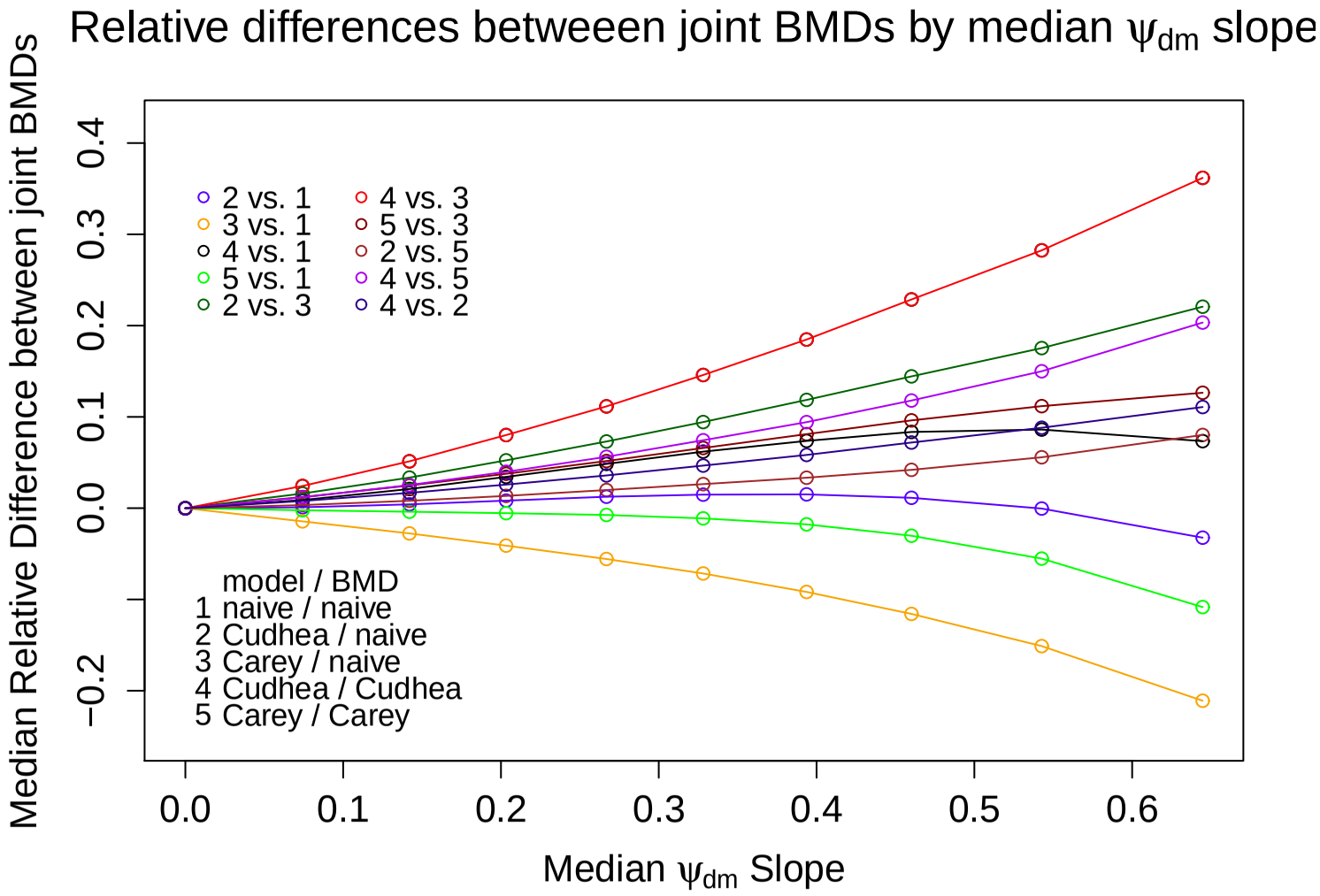

Figure 4.7: The median relative difference between two joint BMD calculation methods vs. median of ψ_3 slope estimate. Because five distinct joint BMD calculation methods are examined, there are a total of 10 different relative differences presented in the plot.

.

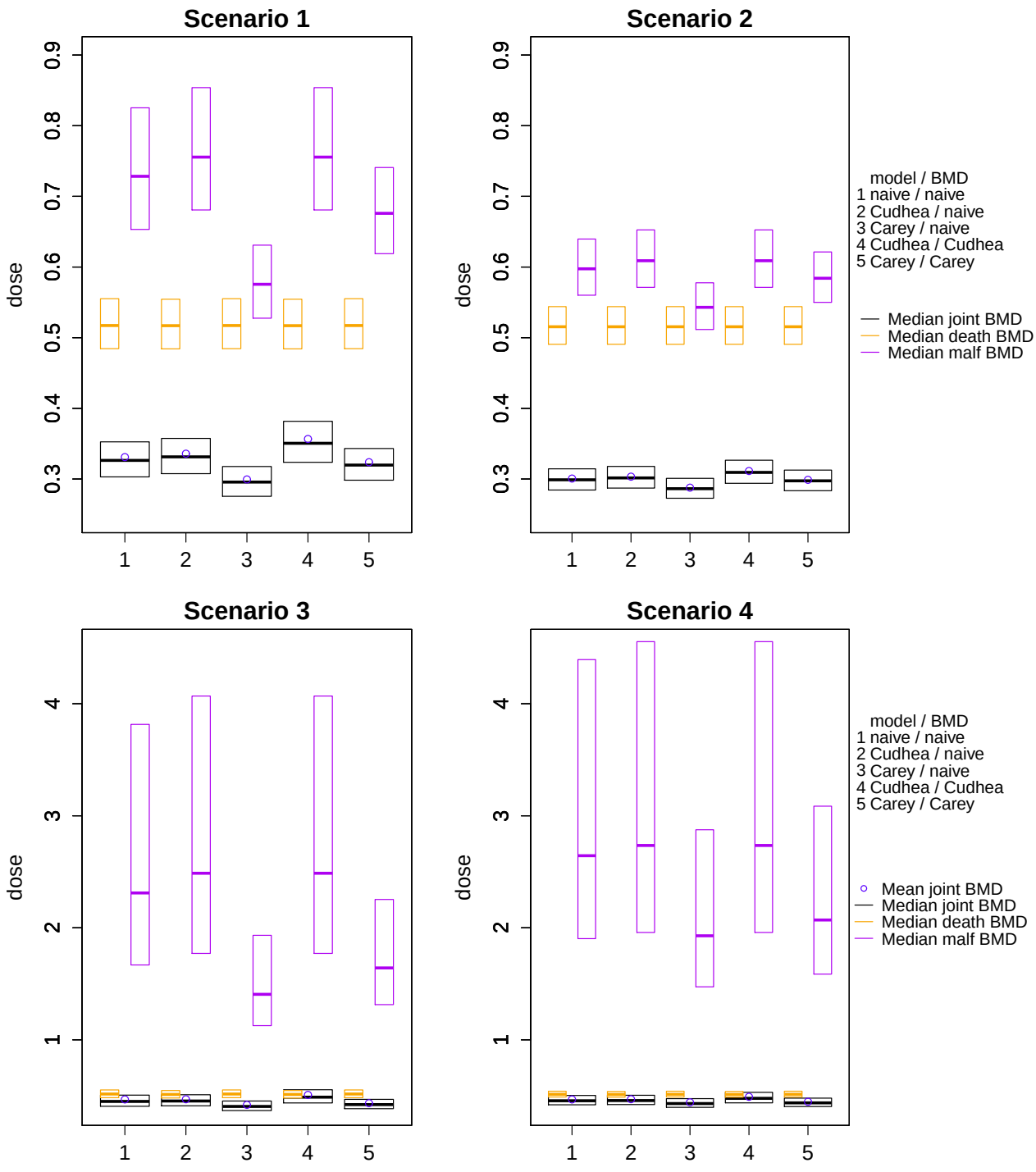

Figure 4.8: Boxplots of joint and individual BMDs for simulation scenarios 1, 2, 3, and 4

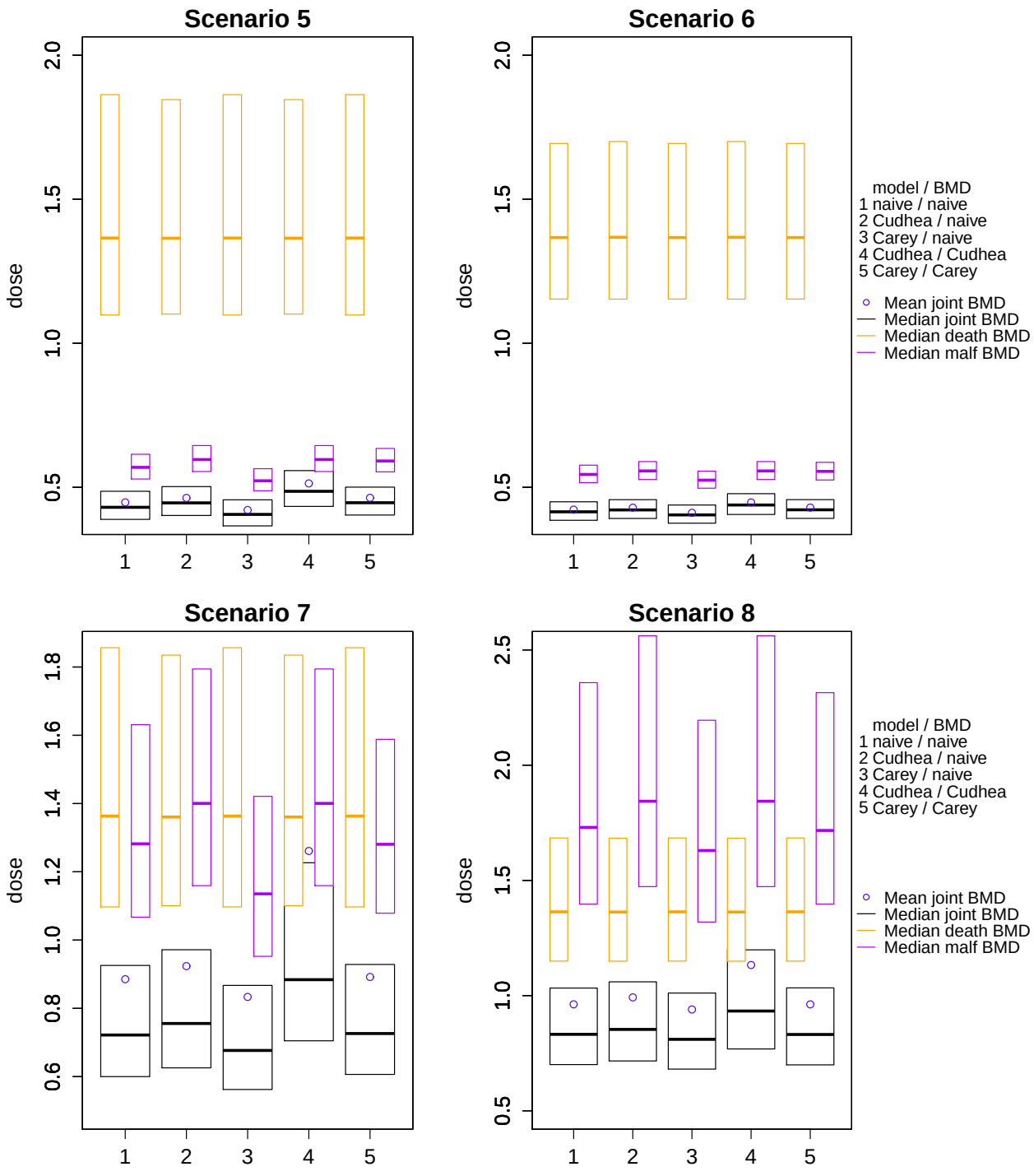

Figure 4.9: Boxplots of joint BMDs for simulation scenarios 5, 6, 7, and 8

4.5.5 Sensitivity to mis-specifying ψ models

It is also of interest to investigate how sensitive the P-D method for joint BMD calculation is to misspecification of the ψ models, and in particular, the ψ_3 model. It has been shown that the p_d and $p_{m|\bar{D}}$ models are robust to ψ model misspecification (Cudhea, 2013). However, for joint BMD calculations, the estimates for the ψ_3 model are also used and over simplifying the model may affect the joint BMD estimates significantly. Figure 4.10 shows the distributions of the joint BMDs, and the corresponding median joint BMDLs (both calculated and empirical) from a simulation study for three different ψ model specifications, for both scenarios 5 and 6 (exact numerical values of summary statistics can be found in appendix). Specifically, the ψ model specifications are as follows:

$$\begin{aligned} ln(\psi_d) &= \alpha_{d_0} + \beta_{m_1} dose \\ ln(\psi_m) &= \alpha_{m_0} + \alpha_{m_1} dose \\ ln(\psi_{dm}) &= \alpha_{dm_0} \end{aligned} \tag{4.5}$$

and

$$\begin{aligned} ln(\psi_d) &= \alpha_{d_0} \\ ln(\psi_m) &= \alpha_{m_0} \\ ln(\psi_{dm}) &= \alpha_{dm_0}. \end{aligned} \tag{4.6}$$

and the original specification (4.4) used in the previous simulations.

We observe that the mean and median of the BMD estimates are slightly higher for models 4.5 and 4.6. This is as expected, since the simplified ψ_{dm} models used should theoretically overestimate the ψ_{dm} parameters at the lower doses, near to where the BMD resides. We also observe that the difference between the BMD distribution for model 4.4 and models 4.5 and 4.6 is greater for scenario 5 than in scenario 6. This is also as expected since the hierarchical correlation is greater for scenario 6. We also note that the gap between estimated BMD and estimated BMDL seems to be relatively stable among all three models, implying that simplifying the second order parameter models does not greatly

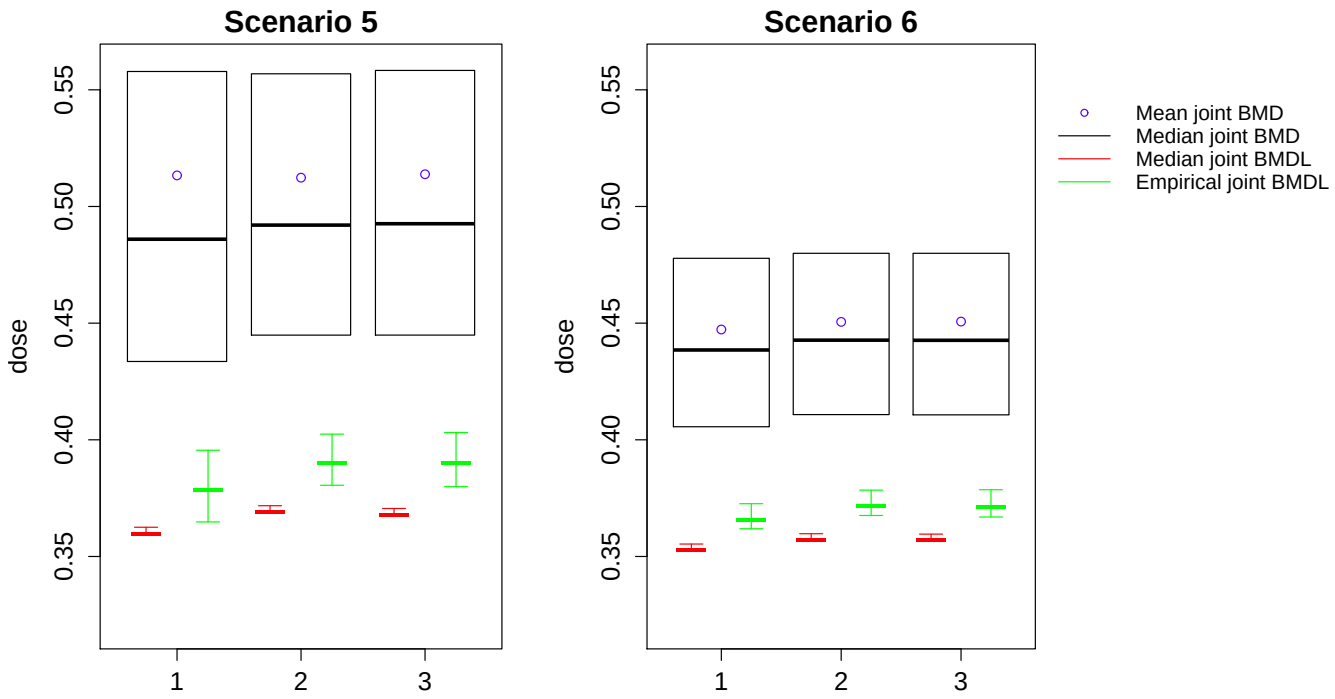

Figure 4.10: Boxplots of joint BMDs for models 4.4, 4.5, and 4.6 for simulation scenarios 5 (left) and 6 (right), and corresponding BMDL estimates.

Table 4.5: Relative differences between median BMD and median BMDL for models 4.4, 4.5, and 4.6 for scenarios 5 and 6

	model 4.4	model 4.5	model 4.6
Scenario 5	0.257	0.247	0.250
Scenario 6	0.192	0.190	0.190

influence BMDL estimation. Table 4.5 shows the median relative difference between BMD and BMDL (defined to be (BMD-BMDL)/BMD) for each method and each scenario examined. Given the difficulty in detecting statistical significance in the slope for second order parameters, if being conservative is a priority, it is perhaps advisable to use the full model to estimate ψ_{dm} even when the dose-response trend for the parameter is not statistically significant.

4.5.6 Bias

While the investigations above have compared how the various methods for BMD calculations differ from one another, they do not provide any insight into bias. In order to investigate bias, we must know the "true" joint risk at each dose so that a true joint BMD can be calculated from the data. How we characterize the true joint risk depends on how we simulate the data. Using Carey's model's as a basis for simulation is a useful approach. The joint distribution of a death and malformation outcome from the same fetus can be expressed as

$$\begin{pmatrix} \tilde{d}_{jk} \\ \tilde{m}_{jk}|\tilde{d}_{jk} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_{m|d} \end{pmatrix}, \begin{pmatrix} \sigma_d & 0 \\ 0 & \sigma_{m|d} \end{pmatrix}\right)$$

where $\mu_d = \tilde{\alpha}_0 + \tilde{\alpha}_1 dose$, $\mu_{m|d} = \tilde{\beta}_0 + \tilde{\beta}_1 dose + \frac{(\rho_{dm}\sigma_m)}{(1+\rho_d(n_k-1))}\left(\frac{\Sigma_{j=1}^{n_k}\tilde{d}_{ij} - n_k(\tilde{\alpha}_0+\tilde{\alpha}_1)dose}{\sigma_d}\right)$, $\sigma_d = 1$ and $\sigma_{m|d} = 1 - \frac{\rho_{dm}n_k}{1+\rho_d(n_k-1)}$. Because the correlation between $\tilde{d}_{jk}$ and $\tilde{m}_{jk}|\tilde{d}_{jk}$ are assumed to be zero in this factorized form, the joint risk formula is given by $1-(1-P(D))(1-P(M|\bar{D}))$. This model is the basis for how the data are simulated in all the simulations presented earlier in the paper, letting μ_d and μ_m equal zero and define cutoffs that deterimine how the latent variable translates into an outcome in practice. Thus, $P(D)$ is defined to be $\Phi(\tilde{\alpha}_0 + \tilde{\alpha}_1 dose|\mu = 0, \sigma = \sigma_d)$ and $P(M|\bar{D})$ is defined to be $\Phi((\tilde{\beta}_0 + \tilde{\beta}_1 dose) + (\rho_{dm}\sigma_m)(1 + \rho_d(n_j-1))^{-1}\left(\frac{\Sigma_{k=1}^{n_j}\tilde{d}_{jk} - n_j(\tilde{\alpha}_0+\tilde{\alpha}_1)dose}{\sigma_d}\right)|\mu = 0, \sigma = \sigma_{m|d})$. However, in the simulation scenarios presented, the cutoffs were determined arbitrarily for each dose. Thus, the exact joint risk is known only for the four doses, making it impossible to calculate a true joint BMD. We conduct a new set of simulations here, where μ_d, μ_m, and ρ_{dm} follow simple linear dose-response trend so that joint risk is known for every dose, and thus the true joint BMD can be calculated for each simulation scenario. Note that the formula for $P(M|\bar{D})$ is dependent on the death outcomes for the entire litter. For a marginal value for $P(M|\bar{D})$, we replace the adjustment covariate with the expected mean of the adjustment covariate, which is 0, so that $P(M|\bar{D}) = \Phi((\tilde{\beta}_0 + \tilde{\beta}_1 dose)|\mu = 0, \sigma = \sigma_{m|d})$. Hierarchical correlation is not ignored in this formula since $\sigma_{m|d}$ is a function of ρ_{dm}.

We examine four simulation scenarios (A, B, C and D). In all four scenarios, the mod-

Table 4.6: ρ_{dm} parameter values for simulation scenarios A, B, and C as well as their corresponding joint risk values by dose

dose	ρ_{dm}				joint risk			
	A	B	C	D	A	B	C	D
0	0	0	0	0.197	0.248	0.248	0.248	0.238
0.75	0	0.0749	0.171	0.268	0.366	0.360	0.353	0.345
1.5	0	0.149	0.332	0.336	0.504	0.489	0.470	0.469
3.0	0	0.291	0.598	0.462	0.771	0.739	0.710	0.722

els for death and malformation cutoff are the same. Specifically, we use the models death cutoff $= -1.25 + 0.2dose$ and malformation cutoff = $-1 + 0.5dose$. These parameters were chosen to reflect a typical study in which the malformation outcome is more sensitive to dose than the death outcome. The ρ_{dm} parameter changes with differs for each scenario. Specifically, in scenario A, we use $g(\rho_{dm}) = 0$, in scenario B, we use $g(\rho_{dm}) = 0.1dose$, in scenario C we use $g(\rho_{dm}) = 0.23dose$, and in scenario D, we use $g(\rho_{dm}) = 0.2 + 0.1dose$ where g is Fisher's z-transformation. In Scenario A, there is no correlation so conditional independence is a valid assumption. In scenario B, the correlation is relatively low even in the higher dose groups. In scenario C, the correlation begins low but increases dramatically with dose while in scenario D, the inter-outcome correlation is relatively high even at lower doses. Table 4.5.6 shows the correlation parameters at each dose as well as the corresponding joint risk based on this model. The true joint BMDs calculated from this joint risk formula for scenarios A, B, C and D are, 0.258, 0.269, 0.312, and 0.318 respectively.

Because the data are simulated under a probit model, and we are assessing bias for these simulations, the joint BMD estimates are also calculated under the probit model. Table 4.5.6 shows the median joint BMD, percent bias, and median joint BMDL for each of the four scenarios using the probit link. We see that the bias is much smaller for scenario A when the probit link is used. For scenarios C and D, in which the hierarchical correlation parameter is the highest, method 4 has the least bias of the five methods examined. For these two scenarios, all other methods substantially underestimate the true joint BMD ranging from 19% to 9.1% bias. Both method 2 and method 4 are substantially less biased

than method 1, which completely ignores hierarchical association. Method 3 is not only the most biased of these methods, but also does not follow an intuitive pattern. Namely, the median BMD does not appear to increase as hierarchical correlation increases. For scenario B, however, method 4 is the worst performing method, overestimating the true joint BMD by 5% while method 1, which assumes conditional independence in both model and joint risk formula, is the most accurate. It is possible that when the hierarchical correlation is low but not zero, the ψ_3 estimates are not as stable, leading to inaccurate estimates. Whatever the reason, the simulations suggest that the methods' accuracy is highly dependent on the strength of the correlation: For high correlation scenarios, method 4 is recommended while for weak correlation scenarios methods 1 and 5 are more reliable.

It is also worth noting that the method 1 results for scenarios B and C appear to confirm that, in scenarios where the hierarchical association is fairly strong, using the simplest method for analyzing the data, and thus completely ignoring conditional dependence, does indeed lead to an underestimation of the joint BMD. Furthermore, at least for scenarios C and D, the more complicated models that do not ignore hierarchical association appear to be significantly more accurate, underlying the importance of using such methods in risk assessment. Methods 2 and 5, while still underestimating the BMD in these scenarios, have median BMD estimates higher than the naive method so seem to account for at least some of the correlation present in the data. It is also worth noting that, differences in median joint BMDL estimates are much smaller compared to the BMD estimates because the variance for the BMD in method 4 is much larger compared to other methods. For example, in scenario C, methods 1, 4 and 5 have almost the same median joint BMDL despite the stark contrast in median joint BMD.

As equally important factor in BMD estimation is the link function used for the death and malformation models. Table 4.5.6 shows the same simulation results as Table 4.5.6 but with the model parameters estimated using a logit link function. For the four scenarios investigated, the BMD results from the logit link are consistently greater than the equivalent results from the probit link. The relative differences between the median logit

link BMDs and and probit link BMDs are shown in the table as well. Median BMD estimates from logit models are 6% to 9% higher than medians from the probit models, exact differences being dependent on scenario and method. The discrepancy between the probit and logit models appears to increase as correlation increases for all five methods, and that discrepancy increases at different rates depending on the method. For example, for method 1, the logit median BMD appears to overestimate the probit median BMD by 6.5% in scenario A and by 6.9% in scenario D. In method 5, the logit median BMD is also 6.5% greater than the probit median BMD in scenario A, but for scenario D it is 8.7% greater. Thus, the simulation results show that using the wrong link function leads to significantly less accurate BMD estimates. In scenario A, where all five methods were fairly accurate using the probit link, the logit link overestimates by about 7%. Similarly, while for scenarios C and D, method 4 was very accurate with the probit models, it overestimates the BMD by about 7.5% using the logit models. This is expected since the data were simulated under a latent multivariate normal framework, which implies the probit link is the correct model.

One of the most important aspects of joint risk assessment is in the low-dose extrapolation that is made possible by fitting dose-response models. Unfortunately, in practice, there is no theoretical basis for assuming one link function is more appropriate than the other. In addition, studies do not typically collect dose-response data for very low doses where the BMD likely resides, making it difficult to ascertain which link function is the most appropriate for a given data set.

4.6 Discussion

In this paper we present five different methods for calculating joint BMDs for developmental toxicity data, four of which are based on models that relax the common but theoretically unsatisfactory conditional independence assumption. These five methods were formally evaluated and compared via simulations under various scenarios. Specifically, the five methods evaluated are as follows:

Table 4.7: Expected BMD, median joint BMDs, and % bias for all five BMD calculations methods examined for simulation scenarios A, B, C, and D (probit link)

		BMD Calculation Method				
	Truth	1	2	3	4	5
Scenario A						
Joint BMD	0.258	0.260	0.260	0.260	0.260	0.260
% bias		-0.508	4.83	4.90	-0.506	-0.542
Joint BMDL		0.245	0.246	0.246	0.234	0.245
Scenario B						
Joint BMD	0.269	0.270	0.274	0.260	0.283	0.270
% bias		-0.158	-1.90	3.55	-5.12	0.430
Joint BMDL		0.245	0.250	0.239	0.239	0.246
Scenario C						
Joint BMD	0.312	0.277	0.288	0.253	0.311	0.280
% bias		11.2	7.52	18.8	0.311	10.0
Joint BMDL		0.243	0.255	0.229	0.245	0.244
Scenario D						
Joint BMD	0.318	0.275	0.289	0.258	0.318	0.288
% bias		13.6	9.10	18.8	0.0286	9.37
Joint BMDL		0.245	0.258	0.235	0.258	0.252

Table 4.8: Expected BMD, median joint BMDs, and % bias for all five BMD calculations methods examined for simulation scenarios A, B, C, and D (logit link)

	Truth	BMD Calculation Method				
		1	2	3	4	5
Scenario A						
Joint BMD	0.258	0.277	0.277	0.277	0.277	0.277
% bias	0	-7.34	-7.35	-0.0736	-0.0735	-0.0737
relative difference between probit and logit BMDs (%)		-6.5	-6.5	-6.5	-6.5	-6.5
Joint BMDL		0.258	0.258	0.257	0.250	0.256
Scenario B						
Joint BMD	0.269	0.288	0.293	0.278	0.303	0.291
% bias	0	-6.88	-8.80	-3.34	-12.5	-7.97
relative difference between probit and logit BMDs (%)		-6.7	-6.9	-6.9	-7.1	-7.8
Joint BMDL		0.258	0.263	0.253	0.257	0.257
Scenario C						
Joint BMD	0.312	0.296	0.308	0.273	0.334	0.305
% bias	0	5.11	1.18	12.4	7.17	2.24
relative difference between probit and logit BMDs (%)		-6.9	-6.9	-7.9	-7.4	-8.9
Joint BMDL		0.256	0.268	0.242	0.252	0.266
Scenario D						
Joint BMD	0.318	0.294	0.310	0.279	0.342	0.313
% bias	0	7.47	2.54	12.4	-7.51	1.57
relative difference between probit and logit BMDs (%)		-6.9	-7.3	-8.1	-7.5	-8.7
Joint BMDL		0.257	0.272	0.247	0.260	0.278

1. Using the naive model and using the naive joint risk formula
2. Using the P-D model and using the naive joint risk formula
3. Using Carey's model and using the naive joint risk formula
4. Using the P-D model and using the P-D formula for joint risk
5. Using Carey's model with mean-adjustment formula for joint risk.

The simulations showed that method 3 consistently gave the most conservative BMD estimates while method 4 was consistently the most anti-conservative. Individual malformation BMDs for the different models showed a similar trend while individual death BMDs tended to be similar. The calculated BMDLs tended to over or underestimate the empirical BMDL (5th percentile of the simulated BMDs) depending on method and scenario. Specifically, the methods that used the naive joint risk formula (1, 2 and 3) tended to have median BMDLs higher than the equivalent empirical BMDs while method 4's median BMDLs were consistently lower than the equivalent empirical BMDs. Therefore, when comparing these methods based on BMDL, method 4 is no longer consistently the most anti-conservative. Also, comparisons between the methods that use the naive joint risk formula and methods that do not suggest that the difference between these methods are indeed dependent on the magnitude of the hierarchical correlation. An investigation in how the specification of the ψ models affects BMD and BMDL estimates confirmed that assuming that ψ_3 does not change with dose increases the BMD and BMDL estimates (assuming the hierarchical correlations increase with dose in reality), but only slightly.

We also conducted a separate simulation study focusing specifically on evaluating differences between these methods as the hierarchical correlation increases. The study showed that the relative differences between the methods, for the most part, increased in a fairly linear and predictable pattern as hierarchical correlation (measured with median ψ_{dm} slope) increased.

Another simulation study was conducted to assess the bias of the five methods. While it is difficult to make any conclusions about methods using the P-D model due to possible lack-of-fit driving the bias estimates, the results suggest that when the hierarchical associations are fairly high, the P-D method gives the most accurate BMD estimates. Both methods 2 and 5 give similar estimates that are higher than those from method 1, as expected, since ignoring this hierarchical correlation should theoretically increase the estimated risk. Method 3 BMD estimates seem to differ very little among the four scenarios, suggesting that the interpretation of β_2 for Carey's model is, indeed, the dose-response for the conditional malformation outcome if no hierarchical correlation exists in a litter. However, since this correlation often does exist, this interpretation has little practical use.

Even as we observe that the joint risk methods based on the P-D model are the most accurate, it must be noted that in the P-D model, every kind of association, between death outcomes and malformation outcomes or otherwise, is conceptualized as between two different fetuses, only existing in a group of implants. Thus, applying this concept to evaluate risk for a single implant is still an ad-hoc approach born out of convenience and intuition. Perhaps because of this, the method can lead to unintuitive results in certain situations. In particular, in the case where we have the marginal $P(D)$ and $P(M|\bar{D})$ increasing relatively slowly with dose, but ψ_{dm} increasing quickly with dose, we may observe that the joint-risk decreases as dose increases. This paradoxical situation is an unfortunate aspect of our formula for joint risk. $F_3(0,0)$ increases as ψ_{dm} increases, but also decreases as p_d and p_m decreases. Thus, in certain scenarios, it is possible for $F_3(0,0)$ to increase, and thus $1 - G_3(0,0)$ to decrease, as dose increases. It should be noted, however, that this scenario where ψ_{dm} increases at a relatively high rate while p_d and $p_{m|\bar{D}}$ do not is an extreme hypothetical scenario that has not been observed in experimental data. To illustrate, consider a scenario where $p_d = .1$, $p_{m|\bar{D}} = .1$, and $\psi_{dm} = 1.1$ for a given dose. Then, according to the P-D method, the joint BMD at this dose is 0.189. In order for the joint BMD to decrease slightly, say to 0.186, then p_d and $p_{m|\bar{D}}$ must increase only slightly, from 0.1 to 0.105 while ψ_{dm} must increase dramatically from 1.1 to 3.0. This illustrates the behavior of the joint risk method but the numerical example is extreme and not expected

Table 4.9: Median adjustment covariate for the EG mice data

dose	0	0.75	1.5	3.0
adjustment covariate (median)	-0.117	-0.492	-0.225	-0.306

Table 4.10: Median adjustment covariate for the 2,4,5-T mice data

dose	0	0.02	0.03	0.045	0.06	0.075	0.09
adjustment covariate (median)	-0.373	-0.402	-0.422	-0.255	-0.929	-0.0141	0.306

in practice.

It should also be noted that method 5, where we use the mean of the adjustment covariate, is still a somewhat simplistic approach that ignores the possibility that this covariate changes with dose. Indeed, assuming the adjustment covariate changes with dose can be thought of as a possible parallel to how the P-D method can assume ψ_{dm} increases with dose since $\beta_2 \left(\frac{\bar{d}_k - logit(\hat{\alpha}_0 + \hat{\alpha}_1 dose)}{\sqrt{logit(\hat{\alpha}_0 - \hat{\alpha}_1 dose)[1 - logit\hat{\alpha}_0 + \hat{\alpha}_1 dose)]/n_k}} \right)$ and $\psi_{dm}(dose)$ both characterize the association between death and malformation outcomes at a particular dose. Thus, fitting a linear regression model to the adjustment covariates and using dose specific means is a potential alternative to using the overall mean. For statistical inference, such an approach may be unsatisfactory since the uncertainty of the linear model will not be accounted for in the resulting BMDL calculations. However, it may offer a more accurate picture of the BMD (and thus the BMDL as well). In both EG and 2,4,5-T data sets, there does not seem to be an obvious pattern between adjustment covariate and dose (Table 4.9 and Table 4.10 shows the median adjustment covariate by dose for the EG study and 2,4,5-T study, respectively), so in practice this alternative approach may not result in significantly different BMD results. How the mean/median adjustment covariate changes with respect to dose in various situations is something that should be studied to assess the appropriateness of this approach (the adjustment covariate may not increase in a linear fashion, for example, making a simple linear regression possibly inappropriate).

The ability to estimate joint BMDs is an essential part of any method that models developmental toxicity data since defining a safe dose is the ultimate goal for these studies. The hierarchical nature of the outcomes of these studies has made if very difficult for a single method to both allow joint BMD estimation and also model death and conditional malformation as unique outcomes with parameters that are easy to interpret. The P-D model and Carey's model are two models that account for hierarchical associations inherent in the data but also model both death and conditional malformation outcomes separately in a relatively straightforward manner, but had no obvious way to integrate the parameters pertaining to the hierarchical association to joint BMD calculation. The joint BMD calculation methods proposed in this paper (methods 4 and 5) are an attempt to improve the utility of these models by making it possible to use them for joint risk estimation, negating the need to fit a completely separate model to the data (such as a model that treats all adverse outcomes as one binary or ordinal variable) in order to answer questions about joint risk.

There are several other avenues of further research to explore. One possibility is to explore the aforementioned extension to method 5, by modeling the adjustment covariate. An investigation of bias conducted in this paper hints that such an approach may not be significantly more accurate in practice, but a more formal assessment may prove useful. Another possibility is to formalize a method for weighting the outcomes (e.g: weighing death as more significant than malformation) in such a way that a weighted joint BMD statistic can be calculated. The models presented open the opportunity for such a method by treating death and malformation as different outcomes. Another is to study existing goodness-of-fit statistics (such as Pearson's Chi-square statistic), or develop new goodness-of-fit tests specific to Carey's model and the P-D model, so that investigators will have better diagnostic tools when considering model fit or choosing which model is most appropriate for the data. It is also of interest to study the distribution of the various BMD calculation methods proposed in yet more detailed simulations, especially scenarios in which the hierarchical correlation is high even in the low doses, and scenarios in which the three correlations increase by dose at different rates.

5.1 Conclusions

Using the methodology developed, we can model the dose-response trends for both death and malformation, as well as three litter-level association parameters, including an association parameter for the hierarchical association between death and malformation. The model allows us to not only relax the potentially erroneous assumption of conditional independence between live and non-live outcomes, and also estimates the association defining this conditional dependence and can model how it changes with dose.

The model assumes that the litter-level correlation can be described by three distinct association parameters, the association between death outcomes, the association between live-outcomes, and the association between death and live outcomes, each with an odds ratio interpretations. This means each association parameter describes the association between two fetal outcomes within a litter, and thus exchangeability is assumed within a litter. The model assumes that each kind of pairing within a litter follows a bivariate Plackett-Dale distribution. The dose-response parameters for death, and then the parameters for malformation and the death-malformation association, are estimated sequentially. The method allows for separate dose-response models for death and malformation (conditional on the fetus not being dead), as well as the three association parameters.

While the model does estimate these association parameters, it does not use a full-likelihood distribution to describe the data. Therefore, the calculation of joint risk is not straightforward. A joint risk formula is developed based on the proposed model that takes advantage of the estimation of the death-malformation association. BMDs calculated using this joint risk formula performed exceptionally well compared to other meth-

ods in simulation scenarios where conditional dependence was strong.

5.2 Advantages

Previous methodology that relaxes the conditional independence assumption either simplify the correlation structure of the data (Christensen's method (Christensen, 2004) uses only one parameter to describe all inherent litter-level correlations) or do not directly estimate all relevant correlation parameters (In Carey's method (Carey, 2006), the parameter for the adjustment term contains information on the magnitude of the hierarchical correlation, but is not a direct estimate). The proposed model, on the other hand, assumes a flexible correlation structure that assumes three separate association parameters to describe all litter-level correlation (like Carey's model) and also allows for the direct estimation (like Christensen's method) for each one. In addition, the method allows for modeling the dose-response for each of these parameters, allowing for a fairly complete picture of the nature of how the data changes as dose changes.

The dose-response parameters also have more intuitive interpretations than the equivalent models in other methods. Christensen's method models cutoffs for a theoretical latent normal distribution. Thus, none of the parameters directly estimates conditional malformation risk (instead, it models death risk and adverse event risk), which is not ideal for toxicologists who are specifically interested in conditional malformation risk. Carey's method models death and conditional malformation in a straightforward manner. However, the malformation model involves an adjustment covariate based on the death-model residuals. Thus, the interpretation of the dose-response parameter is conditional on the adjustment covariate being zero in a litter. This is not ideal since toxicologists are ultimately interested in malformation risk at the population level, not at the litter-specific level (Theoretically, the adjustment variable should be zero on average, but this has not been observed in datasets). The proposed model has an advantage over both latent normal methods in that it models death risk and malformation risk in a straightforward matter, but also does not rely on adjustment covariates to relax conditional in-

dependence. Thus, the dose-response parameter for conditional malformation can be safely interpreted as the population level effect of dose. In addition, the Plackett-Dale framework allows for assuming a Bernoulli random variable for death and conditional malformation rather than a latent normal. This allows for the theoretically justified use of the more widely used logit link function for modeling death and malformation rather than the probit link.

Furthermore, Simulations conducted in chapter 4 suggest the joint risk BMD calculation method proposed for this model is much more accurate than naive methods that assume conditional independence or the ad-hoc non-naive method developed for Carey's method, at least in high correlation scenarios. Given that the ultimate goal of these studies is risk assessment, this is a very promising finding.

5.3 Limitations

The same BMD bias assessment simulations also fond that the proposed method's BMDs overestimated the bias when correlation was low, suggesting that the method is not necessarily appropriate for all data patterns. The second order parameter estimates also tend to have high variances, making any dose-response trend in the association parameters harder to detect in studies with small sample sizes. This is especially problematic for BMD calculations since they rely on the estimate of the hierarchical association parameter. Simulations conducted showed that the BMD calculations are fairly robust to misspecification of the correlation parameter for certain data patterns. However, it is possible this may be of concern for scenarios where the dose-response trend for the association parameter is more extreme.

The high variance of the second order parameters also affect the statistical inference for the BMDs. The BMDL calculations account for uncertainty in the ψ_{dm} parameter. In many data patterns, we see that even as the BMD calculations are the most accurate for the proposed model, they are the least precise and have lower BMDLs than other

methods. Thus, even while we observed that other methods were overly conservative in BMD estimation, in practice, the proposed method can have the lowest BMDL and thus, be the most conservative approach, negating any practical advantage the method may have in terms of accuracy.

A significant limitation of the model, because it considers pairs of fetuses as the unit of data, is its inability to include fetus-level effects. In the context of these toxicity studies, studying the population level effect of dose on the outcomes is a priority so including fetus-level effects are not necessary. However, not being able to include individual-level covariates limits the use of this model outside of this somewhat narrow context.

Finally, extending the Plackett-Dale approach to include other outcomes, namely litter weight, is not as straightforward as methods that assume a latent normal distribution. The Plackett-Dale distribution is well-suited for modeling mixed outcomes. However, the framework we developed would include many more second order parameters to estimate if the method were to be extended to include fetal weight. This, in turn may decrease the precision of the model significantly, and may even lead to a lack of stability. Estimation of the parameters is also likely to be very computationally intensive if fetal weights were to be included. In the proposed method, for each type of pairing, there are at most four possible outcomes, greatly simplifying the computations for parameter estimation. No such shortcut is likely to exist if a continuous outcome is introduced into the model. In addition, as the method is based on bivariate Plackett-Dale distribution, there is no intuitive method for calculating joint risk for three outcomes.

5.4 Future Research

The Plackett-Dale approach we developed to modeling hierarchical outcomes, as yet, does not include fetal weight. An obvious next step would be expand the model to include this outcome. However, as outlined in section 5.3, computational difficulties can be foreseen using this approach, and the resulting model may not be all that pragmatic

to use. However, there is much to be studied concerning the method already developed. More extensive simulations to study the model's behavior, under a wider variety of simulation scenarios, would be useful in understanding in what circumstances, if at all, the model will break down. It would also be of great interest to get a better understanding of which data patterns the BMD estimates are the most accurate in, and which data patterns the BMD estimates tend to be biased in. In particular, the thesis did not explore how the model is affected by varying correlation dose-responses. It would be of interest to see how the model behaves in scenarios, for example, where the hierarchical correlation is low but death and malformation correlation is high, and how that differs from scenarios where all three correlations are high.

Part of the reason this kind of detailed investigation of how changes in specific correlation parameters affect model performance and behavior was not done is that certain combinations of malformation probability and correlation parameter values leads to a correlation matrix for $\tilde{\mathbf{m}}|\tilde{\mathbf{d}}$ that is not positive definite. A more thorough investigation of what combinations are possible for using Carey's latent normal framework to simulate data may act as a useful guide for any future work in the field, especially for conducting simulations for methods not based on a full-likelihood model. A comparison of the data patterns between data simulated from Carey's method and Christensen's method may also be of interest. In particular, how much the single correlation parameter from Christensen's model contributes to conditional dependence has not been studied, and may be of interest to researchers considering using its latent normal framework for simulating data.

Another research path that is potentially of great pragmatic use is to develop simple diagnostic methods for conditional dependence. Carey's model's adjustment term parameter may potentially serve as as a theoretically justified diagnostic statistic. Alternatively, diagnostic plots that are easily interpreted and are informative could be developed. Variations of plotting the distribution of malformation rate and fetal weight against death rate and dose could provide insight for toxicologists who want an intuitive understand-

ing of how much the conditional independence assumption is violated.

Examining different approaches for BMD estimation using Carey's model may also be of interest. The differences between using the sum of the adjustment terms at the individual level and at the litter-level, for example, could be examined. A more formal development of integrating out the adjustment term to obtain a marginal joint risk estimate may also be possible. Since Carey's method is not computationally intensive and easy to implement, and had relative differences between BMDs and BMDLs that were comparable to naive methods, being able to obtain accurate joint BMD estimates with this model would be very useful to toxicologists.

Finally, while this thesis focused entirely on the frequentist approach to the statistical problems present in this data, a Bayesian approach to the problem also shows promise. Indeed, given the multiple layers of hierarchy present in the data, a Bayesian approach may be well suited for the statistical issues present in the data. Exploring a way to model the data that not only explicitly assumes conditional dependence, but also can estimate the correlation parameter defining the conditional dependence, under a Bayesian framework could be worth pursuing.

Appendix A

Supplementary material for chapter 4

A.1 Models fit

For the EG data set, the models fit are as follows:

$$
\begin{aligned}
logit(p_d) &= \beta_{d_0} + \beta_{d_1} dose \\
ln(\psi_d) &= \alpha_{d_0} \\
logit(p_{m|\bar{d}}) &= \beta_{m_0} + \beta_{m_1} dose + \beta_{m_2} dose^2 \\
ln(\psi_m) &= \alpha_{m_0} \\
ln(\psi_{dm}) &= \alpha_{dm_0}
\end{aligned}
$$

The parameter estimates are shown in table A.1.

For the 2,4,5-T data set, the models fit are as follows:

$$
\begin{aligned}
logit(p_d) &= \beta_{d_0} + \beta_{d_1} dose + \beta_{d_2} dose^2 \\
ln(\psi_d) &= \alpha_{d_0} + \alpha_{d_1} dose \\
logit(p_{m|\bar{d}}) &= \beta_{m_0} + \beta_{m_1} dose \\
ln(\psi_m) &= \alpha_{m_0} + \alpha_{m_1} dose \\
ln(\psi_{dm}) &= \alpha_{dm_0}
\end{aligned}
$$

The parameter estimates are shown in table A.1.

Table A.1: Parameter Estimates, Standard Errors, and 95% Confidence Intervals for EG mice data

param	estimate	standard error	95% confidence interval
β_{d_0}	-2.20	0.180	(-2.55, -1.85)
β_{d_1}	0.264	0.101	(0.07, 0.46)
α_{d_0}	0.521	0.139	(0.25, 0.79)
β_{m_0}	-5.26	0.563	(-6.36, -4.16)
β_{m_1}	4.60	0.804	(3.02, 6.18)
β_{m_2}	-0.917	0.219	(-1.35, -0.49)
α_{m_0}	1.23	0.219	(0.80, 1.66)
α_{dm_0}	0.218	0.158	(-0.09, 0.528)

Table A.2: Parameter Estimates, Standard Errors, and 95% Confidenece Intervals for 2,4,5-T data (CD-1 strain)

param	estimate	standard error	95% confidence interval
β_{d_0}	-2.15	0.0551	(-2.26, -2.04)
β_{d_1}	-3.58	3.75	(-10.9, 3.77)
β_{d_2}	304.23	50.69	(204.9, 403.6)
α_{d_0}	0.887	0.149	(0.596, 1.19)
α_{d_1}	16.7	3.01	(10.8, 22.6)
β_{m_0}	-6.33	0.174	(-6.68,-5.99)
β_{m_1}	79.3	3.06	(73.3, 85.3)
α_{m_0}	3.51	0.367	(2.79, 4.23)
α_{m_1}	-18.6	5.94	(-30.2, -6.90)
α_{dm_0}	0.613	0.0739	(0.468, 0.758)

A.2 Summary statistics of adjustment covariates for EG and 2,4,5-T data

For the EG data set, the mean adjustment covariate is -0.0506 (95% confidence interval of (-0.126, 0.0248)) and the median adjustment covariate is -0.306 (95% confidence interval of (-0.311, -0.225)), while the mean for the 2,4,5-T data set is -.0454 (95% confidence interval of (-0.0669, -0.0240)) and the median is -0.509 (95% confidence interval of (-0.509, -0.482)) . Given that the distribution of the adjustment covariates are right-skewed for both distributions, perhaps the median is the more meaningful metric for these cases.

A.3 Parameter values for the simulation scenarios

The cutoff values for death and malformation (c_{d_k} and c_{m_k}, respectively), as well as the between-death correlation (ρ_d), between-malformation correlation (ρ_d), and hierarchical correlation (ρ_{dm}), for the 8 simulation scenarios detailed from section 4.5.1 are shown in Table A.3. The correlation parameters for the 10 simulation scenarios from section 4.5.3 are shown in Table A.3.

A.4 Estimates of mean ψ_{dm} from simulation scenarios

Table A.4 shows the median ψ_{dm} estimates from the model, along with the corresponding dose and corresponding ρ_{dm} from the simulation scenarios presented in section 4.5.6.

A.5 Marginal probabilities for P-D method for joint risk assessment

Using G_3 to describe risk for a single fetus, and that $P(H) = G_3(0, 0)$, it is possible to derive the marginal probalities for each outcome. We know that $P(D) = p_d = (p_d -$

Table A.3: Parameter values for the 8 simulation scenarios

Scenario 1						Scenario 2					
dose	c_{d_k}	c_{m_k}	ρ_d	ρ_m	ρ_{dm}	dose	c_{d_k}	c_{m_k}	ρ_d	ρ_m	ρ_{dm}
0	-1.175	-1.200	0.000	0.000	0.000	0	-1.200	-1.200	0.000	0.000	0.000
0.75	-1.075	-0.900	0.132	0.108	.120	0.75	-0.900	-0.900	0.066	0.054	0.060
1.5	-0.960	-0.590	0.284	0.246	.282	1.5	-0.590	-0.590	0.142	0.123	0.141
3.0	-0.725	0.110	0.600	0.600	.600	3.0	0.110	0.110	0.300	0.300	0.300
Scenario 3						Scenario 4					
dose	c_{d_k}	c_{m_k}	ρ_d	ρ_m	ρ_{dm}	dose	c_{d_k}	c_{m_k}	ρ_d	ρ_m	ρ_{dm}
0	-1.200	-1.200	0.000	0.000	0.000	0	-1.200	-1.175	0.000	0.000	0.000
0.75	-0.900	-1.150	0.132	0.108	.120	0.75	-0.900	-1.075	0.066	0.054	0.060
1.5	-0.590	-0.950	0.284	0.246	.282	1.5	-0.590	-0.960	0.142	0.123	0.141
3.0	0.110	0.400	0.600	0.600	.600	3.0	0.110	-0.725	0.300	0.300	0.300
Scenario 5						Scenario 6					
dose	c_{d_k}	c_{m_k}	ρ_d	ρ_m	ρ_{dm}	dose	c_{d_k}	c_{m_k}	ρ_d	ρ_m	ρ_{dm}
0	-1.175	-1.200	0.000	0.000	0.000	0	-1.175	-1.200	0.000	0.000	0.000
0.75	-1.075	-0.900	0.132	0.108	.120	0.75	-1.075	-0.900	0.066	0.054	0.060
1.5	-0.960	-0.590	0.284	0.246	.282	1.5	-0.960	-0.590	0.142	0.123	0.141
3.0	-0.725	0.110	0.600	0.600	.600	3.0	-0.725	0.110	0.300	0.300	0.300
Scenario 7						Scenario 8					
dose	c_{d_k}	c_{m_k}	ρ_d	ρ_m	ρ_{dm}	dose	c_{d_k}	c_{m_k}	ρ_d	ρ_m	ρ_{dm}
0	-1.175	-1.200	0.000	0.000	0.000	0	-1.200	-1.200	0.000	0.000	0.000
0.75	-1.075	-1.075	0.132	0.108	0.120	0.75	-0.900	-1.150	0.132	0.108	.120
1.5	-0.960	-0.900	0.284	0.246	0.282	1.5	-0.590	-0.950	0.284	0.246	.282
3.0	-0.725	-0.500	0.600	0.600	0.600	3.0	0.110	0.400	0.600	0.600	.600

Table A.4: ρ parameters for 10 simulation scenarios for examining how the difference in methods changes as the hierarchical correlation changes

Scenario 1*				Scenario 2*			
dose	ρ_d	ρ_m	ρ_{dm}	dose	ρ_d	ρ_m	ρ_{dm}
0	0.000	0.000	0.000	0	0.000	0.000	0.000
0.75	0.000	0.000	0.000	0.75	0.0022	0.018	0.020
1.5	0.000	0.000	0.000	1.5	0.0473	0.041	0.047
3.0	0.000	0.000	0.000	3.0	0.100	0.100	0.100
Scenario 3*				Scenario 4*			
dose	ρ_d	ρ_m	ρ_{dm}	dose	ρ_d	ρ_m	ρ_{dm}
0	0.000	0.000	0.000	0	0.000	0.000	0.000
0.75	0.044	0.036	0.040	0.75	0.066	0.054	0.060
1.5	0.0947	0.082	0.094	1.5	0.142	0.123	0.141
3.0	0.200	0.200	0.200	3.0	0.300	0.300	0.300
Scenario 5*				Scenario 6*			
dose	ρ_d	ρ_m	ρ_{dm}	dose	ρ_d	ρ_m	ρ_{dm}
0	0.000	0.000	0.000	0	0.000	0.000	0.000
0.75	0.088	0.072	0.080	0.75	0.110	0.090	0.100
1.5	0.189	0.164	0.188	1.5	0.237	0.205	0.235
3.0	0.400	0.400	0.400	3.0	0.500	0.500	0.500
Scenario 7*				Scenario 8*			
dose	ρ_d	ρ_m	ρ_{dm}	dose	ρ_d	ρ_m	ρ_{dm}
0	0.000	0.000	0.000	0	0.000	0.000	0.000
0.75	0.132	0.108	0.120	0.75	0.154	0.126	0.140
1.5	0.284	0.246	0.282	1.5	0.331	0.287	0.329
3.0	0.600	0.600	0.600	3.0	0.700	0.700	0.700
Scenario 9*				Scenario 10*			
dose	ρ_d	ρ_m	ρ_{dm}	dose	ρ_d	ρ_m	ρ_{dm}
0	0.000	0.000	0.000	0	0.000	0.000	0.000
0.75	0.176	0.144	0.160	0.75	0.198	0.162	0.180
1.5	0.379	0.328	0.376	1.5	0.426	0.369	0.423
3.0	0.800	0.800	0.800	3.0	0.900	0.900	0.900

Table A.5: ρ_{dm} parameter values for simulation scenarios A, B, C, and D, and their corresponding median ψ_{dm} values by dose

dose	ρ_{dm}				median ψ_{dm}			
	A	B	C	D	A	B	C	D
0	0	0	0	0.197	1.002	1.028	1.052	1.848
0.75	0	0.0749	0.171	0.268	1.002	1.360	1.592	2.148
1.5	0	0.149	0.332	0.336	1.001	1.484	2.408	2.496
3.0	0	0.291	0.598	0.462	1.000	2.142	5.509	3.372

Table A.6: Mean, median and stardard deviations for the death BMD and BMDLs, as well as the empirical BMDL values, from all eight simulation scenarios for the P-D and Carey/Naive method

Scenario	Method	mean BMD	median BMD	BMD SD	empirical BMDL
1	P-D	0.523	0.517	0.0557	0.443
	Carey/Naive	0.523	0.517	0.0561	0.443
2	P-D	0.0408	0.515	0.0401	0.458
	Carey/Naive	0.519	0.515	0.0401	0.458
3	P-D	0.517	0.512	0.0517	0.441
	Carey/Naive	0.521	0.516	0.0538	0.443
4	P-D	0.515	0.512	0.0383	0.456
	Carey/Naive	0.517	0.514	0.0389	0.457
5	P-D	2.87	1.36	46.9	0.896
	Carey/Naive	6.93e+11	1.37	4.85e+13	0.892
6	P-D	1.66	1.37	2.94	0.964
	Carey/Naive	1.67	1.37	2.48	0.961
7	P-D	2.86	1.36	47.0	0.895
	Carey/Naive	6.94e+11	1.36	4.86e+13	0.892
8	P-D	1.65	1.36	2.95	0.960
	Carey/Naive	1.66	1.36	2.48	0.960

$F_3(p_d, p_{m|\bar{d}}, \psi_{dm})) + F_3(p_d, p_{m|\bar{d}}, \psi_{dm}) = G_3(m_j|\bar{D}_j = 0, D_{j'} = 1) + G_3(m_j|\bar{D}_j = 1, D_{j'} = 1) = G_3(0,1) + G_3(1,1)$. And since we already assume $P(H) = G_3(0,0)$, this leaves us with $P(M) = G_3(1,0)$. Thus, in our interpretation, the marignal probability for malformation decreases as ψ_{dm} increases.

A.6 Summary statistics for death and malformation BMDs and BMDLs

Table A.6 gives summary statistics for the death BMDs and BMDLs for the 8 simulatoin scenarios presented in section 4.5.1. Table A.6 shows the same summary statistics for malformation BMDs and BMDLs.

Table A.7: Mean, median and standard deviations for the malformation BMD and BMDLs, as well as the empirical BMDL values, from all eight simulation scenarios for all five methods

Scenario	Method	mean BMD	median BMD	BMD SD	empirical BMDL
1	P-D	0.786	0.756	0.161	0.598
	Carey / Naive	0.589	0.576	0.0902	0.475
	Carey / Carey	0.691	0.676	0.107	0.553
	Naive / Naive	0.760	0.728	0.170	0.575
2	P-D	0.616	0.515	0.0401	0.524
	Carey / Naive	0.547	0.543	0.0524	0.471
	Carey / Carey	0.589	0.584	0.0562	0.507
	Naive / Naive	0.604	0.597	0.0635	0.514
3	P-D	9.59	2.49	246.6	1.29
	Carey / Naive	7.16	1.41	348.9	0.892
	Carey / Carey	8.08	1.64	388.5	1.03
	Naive / Naive	5.31	2.31	24.7	1.22
4	P-D	7.64	2.73	56.5	1.39
	Carey / Naive	3.97	1.93	23.1	1.12
	Carey / Carey	4.26	2.07	24.9	1.20
	Naive / Naive	12.3	2.64	185.5	1.36
5	P-D	0.603	0.596	0.0697	0.504
	Carey / Naive	0.528	0.523	0.0574	0.444
	Carey / Carey	0.597	0.591	0.0619	0.506
	Naive / Naive	0.576	0.569	0.0668	0.481
6	P-D	0.560	0.557	0.0478	0.489
	Carey / Naive	0.528	0.525	0.0449	0.460
	Carey / Carey	0.558	0.555	0.0467	0.488
	Naive / Naive	0.548	0.545	0.0462	0.479
7	P-D	1.89	1.40	4.75	0.953
	Carey / Naive	1.51	1.14	8.19	0.791
	Carey / Carey	1.69	1.28	8.97	0.904
	Naive / Naive	1.65	1.28	3.56	0.884
8	P-D	3.61	1.84	31.8	1.15
	Carey / Naive	2.33	1.63	4.59	1.05
	Carey / Carey	2.46	1.72	4.82	1.11
	Naive / Naive	2.63	1.73	10.5	1.10

Table A.8: Mean joint BMD, median joint BMD, median joint BMDL, and empirical joint BMDL for BMD calculation methods 1, 2, and 4, for simulation scenarios 5 and 6

Scenario	Method	Mean BMDL	Median BMD	Median BMDL	Empirical BMDL
5	1	0.513	0.486	0.361	0.380
	2	0.512	0.492	0.371	0.391
	4	0.514	0.493	0.369	0.392
6	1	0.447	0.438	0.354	0.367
	2	0.451	0.442	0.359	0.373
	4	0.451	0.443	0.358	0.373

A.7 Summary statistics of death and malformation BMDs and BMDLs

Table A.7 shows the mean BMD, median BMD, median BMDL, and empirical BMDL for BMD calculation methods 1, 2, and 4, for simulation scenarios 5 and 6.

References

ALTHAM, P. M. E. (1978). Two generalizations of the binomial distribution. *Applied Statistics*, **27** 162–167.

CAREY, A. (2006). *Dose-Response Models for Mixed Dependent Outcomes in Developmental Toxicity*. Ph.D. thesis, Harvard University.

CARR, G. J. and PORTIER, C. J. (1991). An evaluation of the rai and van ryzin dose-response model in teratology. *Risk Analysis*, **11** 111–120.

CATALANO, P. J. and RYAN, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, **87** 651–658.

CATALANO, P. J. and RYAN, L. M. (1994). *Chemical Risk Assessment and Occupational Health: Current Applications, Limitations, and Future Prospects*, chap. 12. Greenwood Publishing Group, Westport, CT, 123–136.

CATALANO, P. J., RYAN, L. M. and SCHARFSTEIN, D. O. (1994). Modeling fetal death and malformation in developmental toxicity. *Risk Analysis*, **14** 611–619.

CATALANO, P. J., SCHARFSTEIN, D. O., RYAN, L. M., KIMMEL, C. A. and KIMMEL, G. L. (1993). Statistical model for fetal death, fetal weight, and malformation in developmental toxicity studies. *Teratology*, **47** 281–290.

CHEN, J. (1993). A malformation incidence dose-response model incorporating fetal weight and/or litter size as covariates. *Risk Analysis*, **13** 559–564.

CHEN, J. and GAYLOR, D. W. (1992). Correlations of developmental end points observed after 2,4,5-trichlorophenoxyacetic acid exposure in mice. *Teratology*, **45** 241–246.

CHEN, J. J., KODELL, R. L., HOWE, R. B. and GAYLOR, D. W. (1991). Analysis of trinomial respones from reproductive and developmental toxicity experiments. *Biometrics*, **47** 1049–1058.

CHRISTENSEN, J. C. (2004). *Likelihood Methods for Clustered Discrete and Continuous Outcomes in Developmental Toxicology*. Ph.D. thesis, Harvard University.

CRUMP, K. S. (1984). A new method for determining allowable daily intakes. *Fundamental and Applied Toxicology*, **4** 854–871.

CUDHEA, F. P. (2013). *A Novel Method for Modeling Hierarchical Outcomes in Developmental Toxicity Data Based on the Plackett-Dale Distribution*. Ph.D. thesis, Harvard University.

DALE, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42** 909–917.

FITZMAURICE, G. M. and LAIRD, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, **90** 845–852.

GAYLOR, D., RYAN, L., KREWSKI, D. and ZHU, Y. (1998). Procedures for calculating benchmark doses for health risk assessment. *Regulatory Toxicology and Pharmacology*, **28** 150–164.

GEYS, H., M., R. M., CATALANO, P. J. and MOLENBERGHS, G. (2001). Two latent variable risk assessment approaches for mixed continuous and discrete outcomes from developmental toxicity data. *Journal of Agricultural, Biological, and Environmental Statistics*, **6** 340–355.

KIMMEL, C. A. and GAYLOR, D. (1988). Issues in qualitative and quantitative risk analysis for developmental toxicology. *Risk Analysis*, **8** 15–20.

KIMMEL, C. A. and PRICE, C. J. (1990). *Handbook of In Vivo Toxicity Testing*, chap. 12. Academic Press. Inc., 271–300.

KUPPER, L. L. and HASEMAN, J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics*, **34** 69–76.

LEFKPOULOU, M., MOORE, D. and RYAN, L. (1989). The analysis of multiple correlated binary outcomes: Application to rodent teratology experiments. *Journal of the American Statistical Association*, **484** 163–174.

LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73** 13–22.

MENDELL, N. R. and ELSTON, R. C. (1974). Multifactorial qualitative traits: Genetic analysis and prediction of recurrence risks. *Biometrics*, **30** 41–57.

MOLENBERGHS, G., GEYS, H. and BUYSE, M. (2001). Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine*, **20** 3023–3038.

MOLENBERGHS, G. and LESAFFRE, E. (1994). Marginal modeling of correlated ordinal data using a multivariate plackett distribution. *Journal of the American Statistical Association*, **89** 633–644.

OCHI, Y. and PRENTICE, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika*, **71** 531–543.

PLACKETT, R. (1965). A class of bivariate distributions. *Journal of the American Statistical Association*, **60** 516–522.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. and FLANNERY, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*, chap. 9. Cambridge University Press, 442–486.

PRICE, C. J., KIMMEL, C. A., TYL, R. W. and MARR, M. C. (1985). The developmental toxicity of ethylene glycol in rats and mice. *Toxicology and Applied Pharmacology*, **81** 113–127.

RAI, K. and VAN RYZIN, J. (1985). A dose-response model for teratological experiments involving quantal responses. *Biometrics*, **41** 1–9.

REGAN, M. M. and CATALANO, P. J. (1999). Bivariate dose-response modeling and risk estimation in developmental toxicology. *Journal of Agricultural, Biological, and Environmental Statistics*, **4** 217–237.

RYAN, L. (1992). Quantitative risk assessment for developmental toxicity. *Biometrics*, **48** 163–174.

RYAN, L. M., CATALANO, P. J., KIMMEL, C. A. and KIMMEL, G. L. (1991). Relationship between fetal weight and malformation in developmental toxicity studies. *Teratology*, **44** 215–223.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY, H. (1991). Guidelines for developmental toxicity risk assessment. *Federal Register*, **56** 63798–63826.

WEDDERBURN, W. M., ROBERT (1974). Quasilikelihood functions, generalized linear models and the gauss-newton method. *Biometrika*, **61** 439–447.

WILLIAMS, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, **31** 949–952.